FINAL REPORT

OPTIMIZATION, ESTIMATION, AND CONTROL OF INLET CONTROL SYSTEMS FOR AEROSPACE VEHICLES

G. Allgaier
R. Stefani
S. Yakowitz

July, 1970

Prepared under Contract NGR-03-002-115

ENGINEERING EXPERIMENT STATION
COLLEGE OF ENGINEERING
THE UNIVERSITY OF ARIZONA
TUCSON, ARIZONA

FINAL REPORT

Optimization, Estimation and Control of Inlet Control Systems for Aerospace Vehicles

G. Allgaier
R. Stefani
S. Yakowitz

July, 1970

Distribution of this report is provided in the interest of information exchange. Responsibility for the contents resides in the authors' organization that prepared it.

Prepared under Contract NGR-03-002-115

by

The University of Arizona
Department of Electrical Engineering
Tucson, Arizona

for

Lewis Research Center
National Aeronautics and Space Administration

TABLE OF CONTENTS

PREFACE

The following report is divided into four parts. The first is a discussion of the application of Kalman filter theory to systems with time delay. Application to the "inlet problem" is included in the appendix. The nonlinear Kalman equations were solved by converting to a set of discrete equations (as discussed by Meditch). Further work is planned which will develop a more general computational procedure for more than one time delay. Also the problem with a large variety of measurements and states is being attacked.

The next problem is similar but the noise and time delay are not included in the model. An appropriate feedback controller is described by a set of parameters which specify the gain, pole, and zero locations. These parameters are varied to determine the combination of parameter values which minimize the unstart frequency. This was done by plotting the number of unstarts as a function of two of the parameters. The stability of the system was facilitated by having a pole-zero excess of two which requires that the feedback compensator have more zeros than poles. The requisite number of poles could be added to when the compensator is actually built at an appropriately high frequency.

The next section includes a discussion of estimation of the output matrix when noisy measurements are made of both the states and the output. Normal least squares estimation results in an estimate which has a fixed error even for a large number of measurements. This bias can be removed by modifying the objective function and then linearizing the resulting equations for the estimates of the output matrix. The resulting estimate is not only linear but it also is iterative. That is, a new estimate can be made as more measurements are taken without having to store all the past values of the state and output.

The final section is a detailed mathematical investigation of the properties of a uniform random search for the maximum of a function of several parameters. As the final page points out, in the absence of some regularity to the function in the search region only a very close inspection of the function will suffice. In this most difficult situation a random search will do the job. How well it can do the job is thoroughly explored in both the case of exact and noisy measurements of the function.

T. L. Williams

OPTIMAL CONTROL OF LINEAR SYSTEMS WITH TIME DELAY, PLANT NOISE AND OBSERVATION NOISE

The following paragraphs develop an optimal control for a linear system whose measured output is a delayed linear combination of the system states under a quadratic performance index. It includes the effects of plant and observation noise. The optimal control is generated by a cascade combination of a Kalman filter, a linear predictor and an optimal controller. The basis for much of this work was done by Kleinman.[1]

It is assumed that the plant is time-invariant and may be expressed by:

$$\dot{x}(t) = Fx(t) + Cu(t) + \omega(t) \tag{1}$$

$$z(t) = Hx(t - \tau) + v(t - \tau) \tag{2}$$

where $\omega(t)$ is the plant noise and $v(t)$ is the measurement noise with the following autocorrelation functions

$$E[\omega(t)\omega'(\sigma)] = Q\delta(t - \sigma) \tag{3}$$

$$E[v(t)v'(\sigma)] = R\delta(t - \sigma) \tag{4}$$

The noise $\omega(\tau)$ is presumed statistically independent from the noise $v(t)$. The system is shown in block diagram form in Figure 1. The optimal control of the plant is achieved in three steps.

1. The <u>optimal estimate</u>: $\hat{x}(t - \tau)$ from $z(t)$
2. The <u>optimal prediction</u>: $\hat{x}(t)$ from $\hat{x}(t - \tau)$
3. The <u>optimal control</u>: $u(t)$ from $\hat{x}(t)$

There will be some feedback between the blocks in Figure 2, but the diagram shows the essential form of the solution.

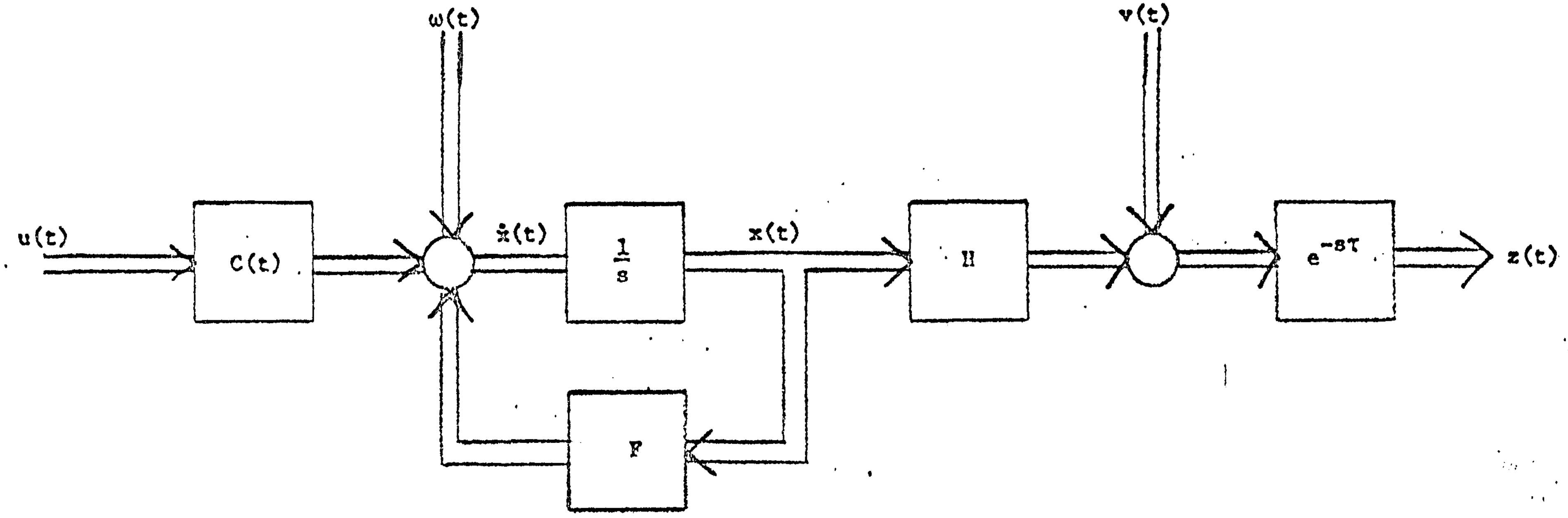


Figure 1. Block Diagram of Plant to be Controlled

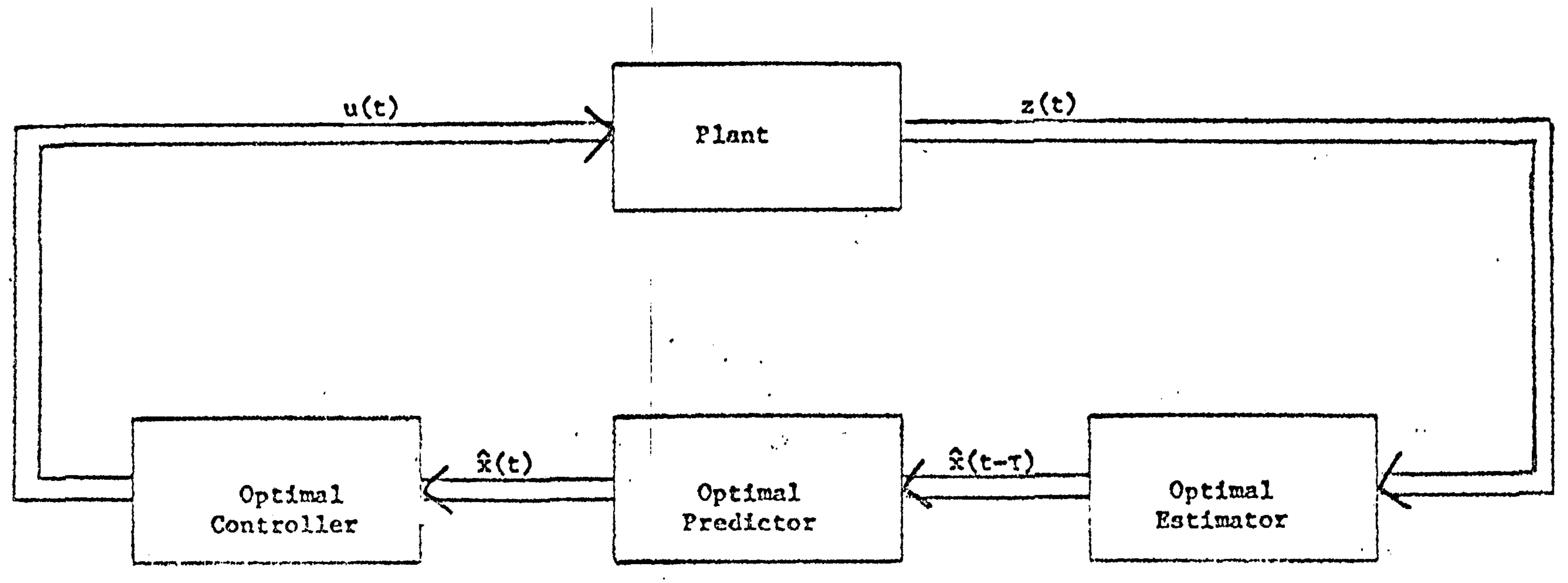


Figure 2. Generalized Block Diagram of Optimal Controller to be Developed

The solution is achieved by considering plants with the following characteristics:

1. Linear system with output $z(t) = x(t)$ - (This yields the optimal controller in Figure 2).

2. Linear system with delayed output $z(t) = x(t - \tau)$. (This yields the optimal predictor with the same controller achieved in (1)).

3. Linear system with delayed output and measurement noise $z(t) = Hx(t - \tau) + v(t - \tau)$. This yields the optimal estimator with the same predictor as (2) and the same controller as (1)).

I. Linear System with Output $z(t) = x(t)$:

With the plant equations:

$$\dot{x}(t) = Fx(t) + Cu(t) + \omega(t) \tag{5}$$

$$z(t) = x(t) \tag{6}$$

It is desired to minimize the quadratic cost functional

$$J(u) = E\{ \lim_{T\to\infty} \frac{1}{T} \int_0^T [x(+)'Ax(t) + u'(t)Bu(t)]dt\} \tag{7}$$

The solution to this problem is well-known[2]

$$u^*(t) = S(t)\hat{x}(t|t) \tag{8}$$

where
$$S(t) = -B^{-1}C'W(t) \tag{9}$$

where $W(t)$ is the solution to the Matrix-Ricatti equation

$$\dot{W}(t) = -F'W(t) - W(t)F + W(t)CB^{-1}C'W(t) - A \tag{10}$$

The value of the performance index is

$$J(u*) = \text{trace}\{WQ\} \tag{11}$$

The solution is shown in Figure 3.

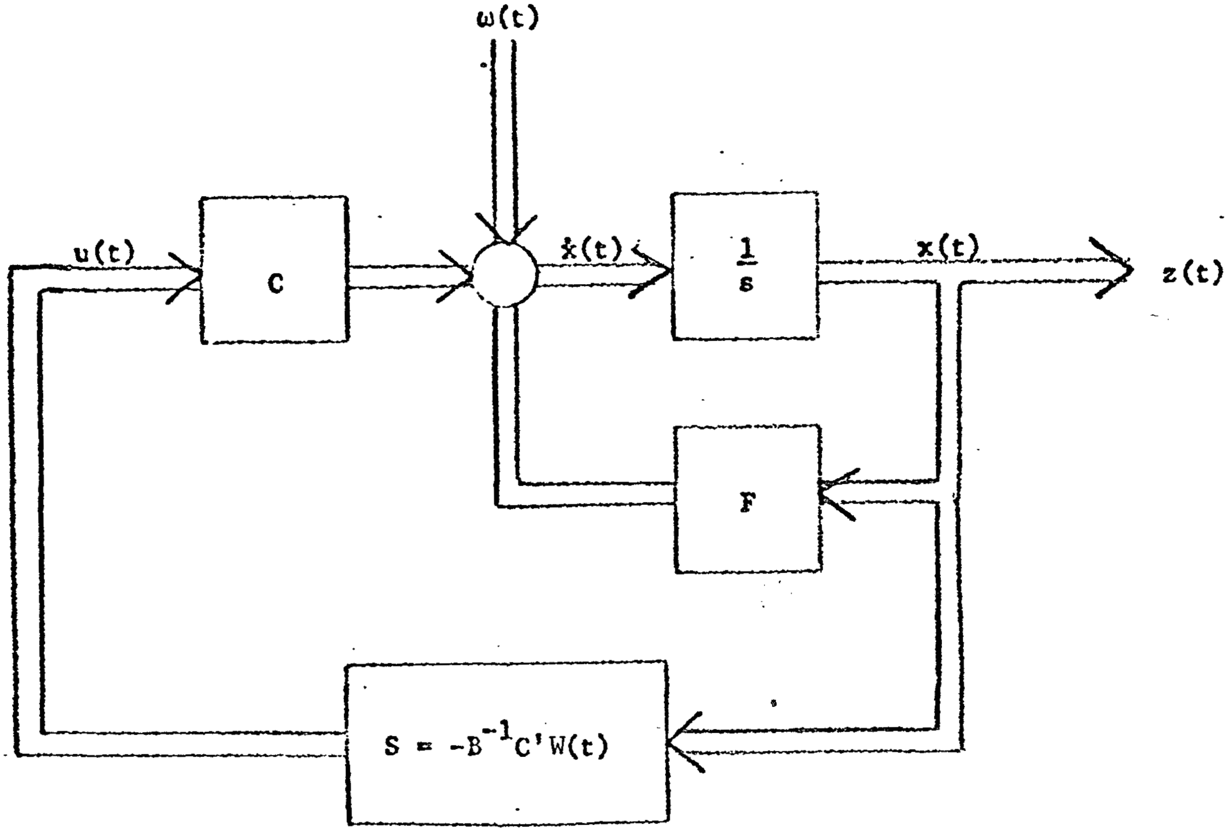


Figure 3. Optimal Control of Plant with no Time Delay and no Measurement Noise

II. Linear System with Output $z(t) = x(t - \tau)$

The plant and output equations are now

$$\dot{x}(t) = Fx(t) + Cu(t) + \omega(t) \tag{12}$$

$$z(t) = x(t - \tau) \tag{13}$$

It is desired to determine the optimal control u(t) to minimize the same quadratic cost functional (7) as above. This is done by investigating the prediction process:

$$\hat{x}(t) = E\{x(t) | z(\sigma), \sigma \leq t\} \tag{14}$$

In order to generate $\hat{x}(t)$, note that since

$$z(t) = x(t - \tau)$$

that
$$\dot{z}(t) = \dot{x}(t - \tau) \tag{15}$$

or
$$\dot{x}(t) = \dot{z}(t + \tau) \tag{16}$$

Substituting in the equations (12) and (13):

$$\dot{x}(t) = Fx(t) + Cu(t) + \omega(t)$$

$$\begin{aligned} \dot{z}(t) &= \dot{x}(t - \tau) \\ &= Fx(t - \tau) + Cu(t - \tau) + \omega(t - \tau) \end{aligned}$$

or,
$$\dot{z}(t) = Fz(t) + Cu(t - \tau) + \omega(t - \tau) \tag{17}$$

Since the control input u(t) is a deterministic process, and since the system is linear, we define $z_u(t)$ to be the contribution of z(t) due to u(t).

$$z(t) = z_u(t) + r(t) \tag{18}$$

where r(t) is the contribution of z(t) due to noise. (17) may then be rewritten

$$\dot{z}(t) = F[zu(t) + r(t)] + Cu(t - \tau) + \omega(t - \tau) \tag{19}$$

(19) may, in turn, be written as two independent equations

$$\dot{z}_u(t) = Fz_u(t) + Cu(t - \tau) \tag{20}$$

$$\dot{r}(t) = Fr(t) + \omega(t - \tau) \tag{21}$$

where (20) relates deterministic inputs and outputs and (21) relates noisy inputs and outputs. This separation is possible because of the linear system. From (14),

$$\hat{x}(t) = E\{x(t)|z(\sigma), \sigma \leq t\}$$

since $$x(t) = z(t + \tau) = z_u(t + \tau) + r(t + \tau)$$

then
$$\begin{aligned} \hat{x}(t) &= E\{z_u(t + \tau) + r(t + \tau)|z(\sigma), \sigma \leq t\} \\ &= E\{z_u(t + \tau)|z(\sigma), \sigma \leq t\} + E\{r(t + \tau)|z(\sigma), \sigma \leq t\} \\ &= z_u(t + \tau) + E\{r(t + \tau)|r(\sigma), \sigma \leq t\} \qquad (22) \end{aligned}$$

The second term of (22) becomes

$$E\{r(t + \tau)| r(\sigma), \sigma \leq t\} = e^{F\tau}r(t)$$

since (t) is white noise.[3] Therefore (22) becomes

$$\hat{x}(t) = z_u(t + \tau) + e^{F\tau}r(t) \qquad (23)$$

From the definition of x(t) given in (14),

$$\hat{x}(t) = E\{x(t)|z(\sigma), \sigma \leq t\} \qquad (14)$$

we have generated $\hat{x}(t)$, the <u>least mean-squared error</u> prediction of x(t). The implementation of (23) is shown in Figure 4. As indicated in Figure 4, the optimal controller remains to be determined. First, we develop the appropriate system equations. Taking the derivative of both sides of (23) results in

$$\dot{\hat{x}}(t) = \dot{z}_u(t + \tau) + e^{F\tau}\dot{r}(t) \qquad (24)$$

Substituting from (20) and (21) in (24):

$$\dot{\hat{x}}(t) = Fz_u(t + \tau) + Cu(t) + e^{F\tau}[Fr(t) + \omega(t - \tau)] \qquad (25)$$

Substituting from (23) for $z_u(t + \tau)$,

$$\dot{\hat{x}}(t) = Fx(t) - Fe^{F\tau}r(t) + Cu(t) + e^{F\tau}Fr(t) + e^{F\tau}\omega(t - \tau) \qquad (26)$$

Nothing that $Fe^{F\tau} = e^{F\tau}F$, (26) becomes

$$\dot{\hat{x}}(t) = F\hat{x}(t) + Cu(t) + e^{F\tau}\omega(t - \tau) \qquad (27)$$

The optimal controller for the system expressed by (27) is determined by the minimization of the quadratic cost functional (7).

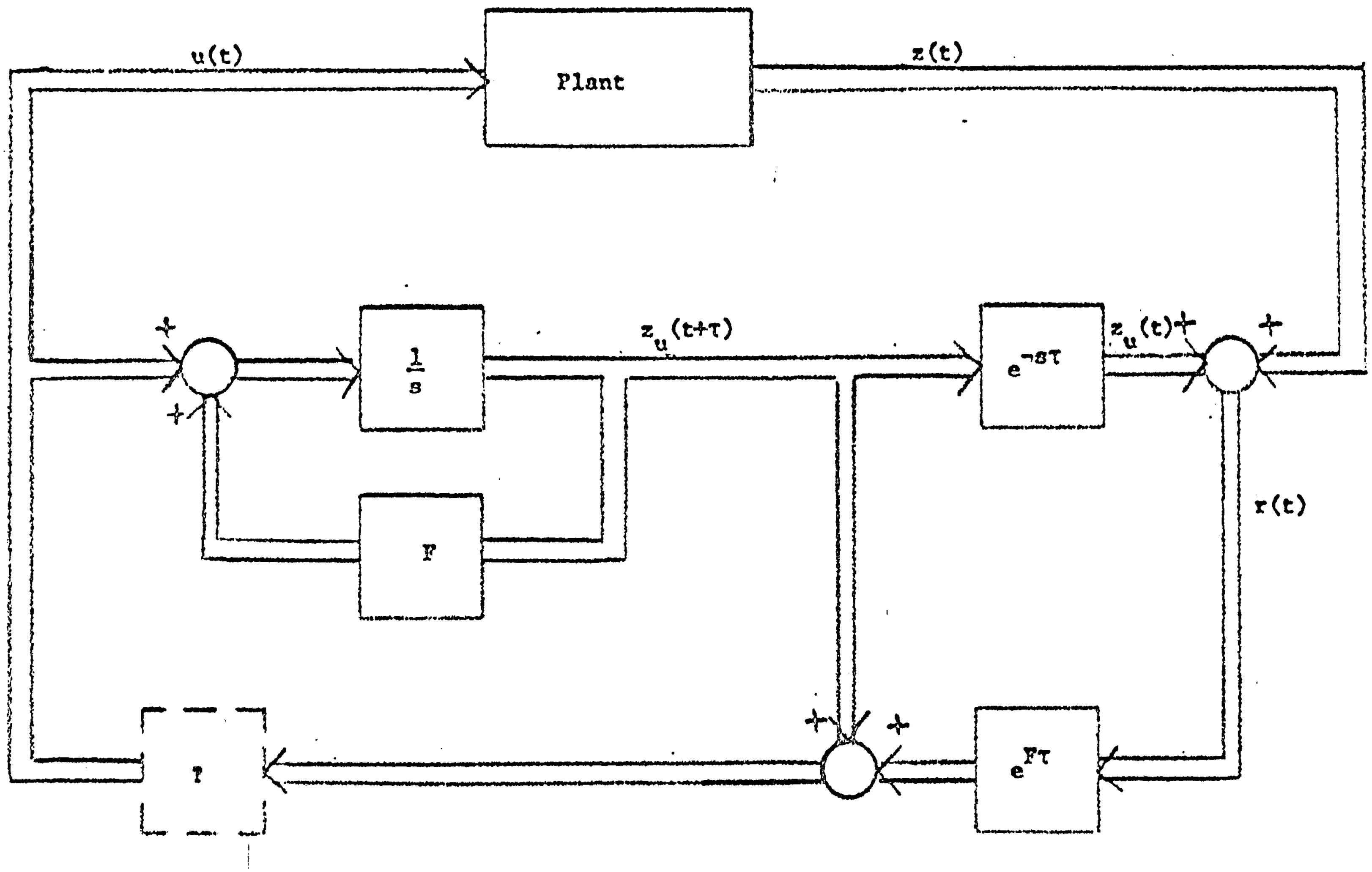


Figure 4. Optimal Predictor of x(t) Given x(t-τ)

First, it is noted that

$$\begin{aligned} E\{x'(t)Ax(t)\} &= E\{[\hat{x}(t) + e(t)]'A[\hat{x}(t) + e(t)]\} \\ &= E\{\hat{x}(t)'A\hat{x}(t)\} + E\{\hat{x}(t)'Ae(t)\} \\ &\quad + E\{e(t)'A\hat{x}(t)\} + E\{e(t)'Ae(t)\} \\ &= E\{\hat{x}(t)'A\hat{x}(t)\} + E\{e(t)'Ae(t)\} \end{aligned} \tag{28}$$

Since $E\{\hat{x}(t)'e(t)\} = E\{e(t)'x(t)\} = 0$ when $\hat{x}(t)$ is the least mean square error estimate of $x(t)$. Hence $J(u)$ can be written (assuming interchange of $\lim\{\cdot\}$ and $E\{\cdot\}$ operators).

$$J(u) = \lim_{T\to\infty} [\frac{1}{T} \; E\{\int_0^T [e(t)'Ae(t) + \hat{x}(t)'A\hat{x}(t) + u(t)'Bu(t)]dt\}] \tag{29}$$

Since only the last two terms depend on u, (the first term is minimized by the prediction process) it is only necessary to minimize

$$J_1(u) = \lim_{T\to\infty} [\frac{1}{T} \{E \int_0^T [\hat{x}(t)'A\hat{x}(t) + u(t)'Bu(t)]dt\}] \tag{30}$$

where $\hat{x}(t)$ is generated by: (27)

$$\dot{\hat{x}}(t) = F\hat{x}(t) + Cu(t) + e^{F\tau}\omega(t - \tau) \tag{27}$$

We note that (27) is of the same form as the "no delay" plant equation (5)

$$\dot{x}(t) = Fx(t) + Cu(t) + \omega(t) \tag{5}$$

except that $\omega(t)$ is replaced by $e^{F\tau}\omega(t - \tau)$. The optimal control solution is therefore of the same form as that generated by (8) and (9).

$$u(t) = S(t)\hat{x}(t|t) \tag{8}$$

$$S(t) = -B^{-1}C'W(t) \tag{9}$$

Note, however, that the value of the Performance Index is no longer the same. In fact, it is dependent on the time delay and can be shown to be[1]

$$J(u^*) = \text{trace } \{A \int_0^\tau e^{F\sigma}Qe^{F'\sigma}d\sigma\} + \text{trace } \{We^{F\tau}Qe^{F'\tau}\} \tag{28}$$

III. The Noise-Time Delay Problem

Consider now the system described by (29) and (30)

$$\dot{x}(t - \tau) = Fx(t - \tau) + Cu(t - \tau) + \omega(t - \tau) \tag{29}$$

$$z(t) = Hx(t - \tau) + v(t - \tau) \tag{30}$$

Let $$\hat{x}(t - \tau) = E\{x(t - \tau)|z(\sigma),\ \sigma \leq t\} \tag{31}$$

be the least mean-squared estimate of $x(t - \tau)$ based on the observation of $z(\sigma)$, $\sigma \leq t$.

The solution to this problem is well-known[4] and is a slight modification of the Kalman filter which includes the effect of a deterministic control $u(t)$. Solving (32) for $P(t)$, the error covariance matrix

$$\dot{P}(t) = FP(t) + P(t)R' - P(t)H'R^{-1}HP(t) + Q \tag{32}$$

The solution for $\hat{x}(t - \tau)$ is given by

$$\dot{\hat{x}}(t - \tau) = Fx(t - \tau) + Cu(t - \tau) + P(t - \tau)H'R^{-1}[z(t) - H\hat{x}(t - \tau)] \tag{33}$$

The implementation is shown in Figure 5, where $\overline{P}$ is the steady-state solution of (32).

To determine $u(t)$, consider the quadratic cost functional (7) modified to include the effect of the time delay

$$J(u) = \lim_{T\to\infty} \frac{1}{T} E\{ \int_{\tau}^{T} [x'(t - \tau)Ax(t - \tau) + u'(t - \tau)Bu(t - \tau)]dt\} \tag{34}$$

As before, letting $x(t) = \hat{x}(t) + e(t)$, (34) becomes

$$J(u) = \lim_{T\to\infty} \frac{1}{T} \{E \int_{\tau}^{T} [e'(t - \tau)Ae(t - \tau) + \hat{x}(t - \tau)'A\hat{x}(t - \tau)$$

$$+ u'(t - \tau)Bu(t - \tau)]dt\} \tag{35}$$

Once again, $e(t)$ is independent of $u(t)$ and for all t, $E\{e'(t)Ae(t)\}$ is at an absolute minimum and it can be shown[1] that

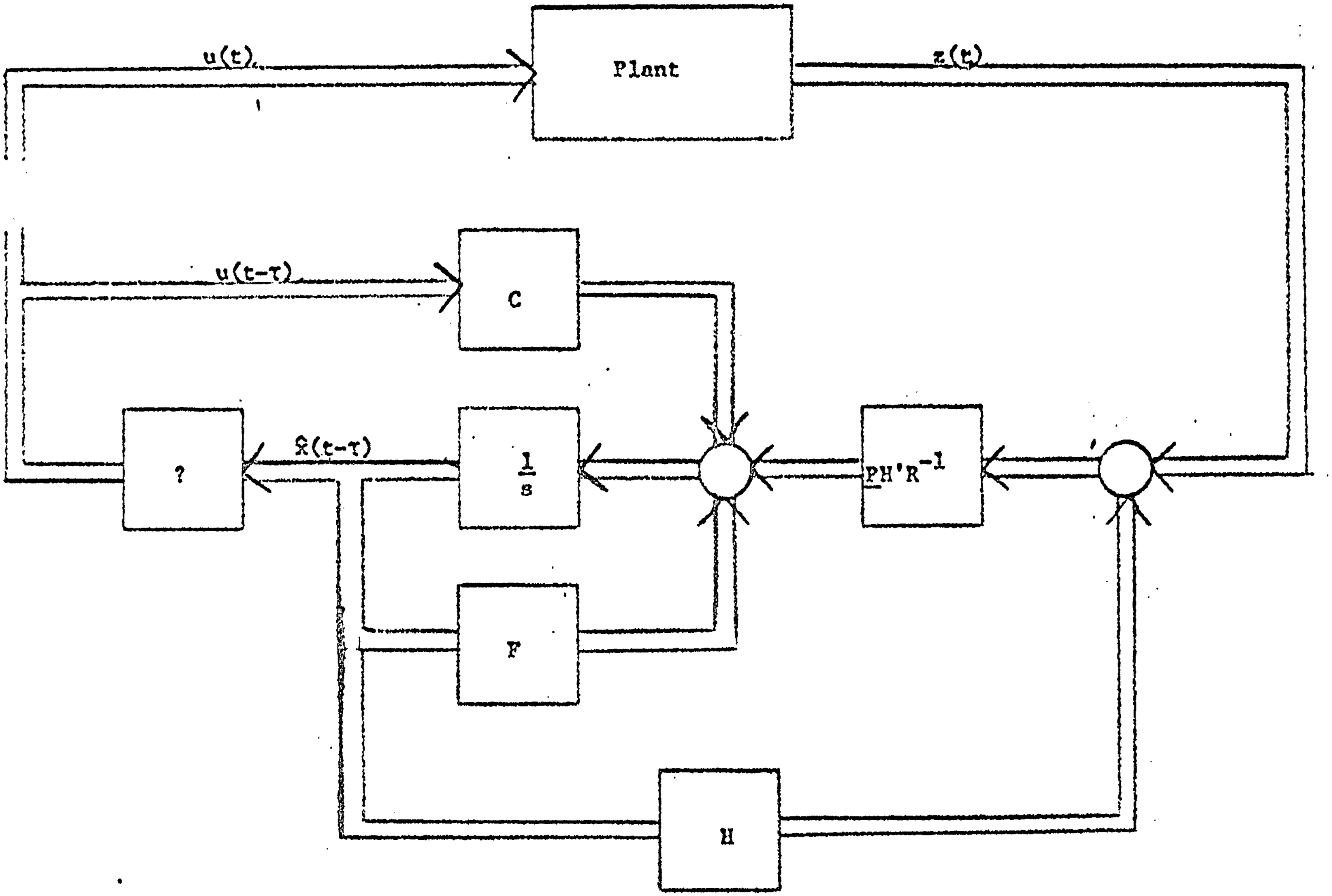


Figure 5. Optimal Estimate of $\hat{x}(t-\tau)$

$$\lim_{T\to\infty} \frac{1}{T} E\{ \int_{\tau}^{T} e'(t-\tau)Ae(t-\tau)dt\} = \text{trace}(A\bar{P})$$

Since (35) is evaluated in the limit as T goes to ∞ it may be expressed equivalently as,

$$\lim_{T\to\infty} \frac{1}{T} E\{ \int_{\tau}^{T} [\hat{x}(t)'A\hat{x}(t) + u(t)'Bu(t)]dt\} \qquad (36)$$

Substituting (30) in (33) for $z(t)$

$$\dot{\hat{x}}(t-\tau) = F\hat{x}(t-\tau) + Cu(t-\tau) + \bar{P}(t-\tau)H'R^{-1}[Hx(t-\tau) + v(t-\tau) - H\hat{x}(t-\tau)] \qquad (37)$$

and that $e(t) = x(t) - \hat{x}(t)$ and using the steady state value of $P(t)$

$$\dot{\hat{x}}(t) = F\hat{x}(t) + Cu(t) + \bar{P}H'R^{-1}[He(t) + v(t)] \qquad (38)$$

Wonham has shown [5] that the process

$$PH'R^{-1}[He(t) + v(t)] \qquad (39)$$

is a white noise process with covariance matrix $\tilde{Q}$

$$\tilde{Q}\delta(t-\sigma) = E[\tilde{q}(t)\tilde{q}(t)'] = PH'R^{-1}HP\delta(t-\sigma) \qquad (40)$$

Therefore (38) can be written

$$\dot{\hat{x}}(t) = F\hat{x}(t) + Cu(t) + \hat{q}(t) \qquad (41)$$

Recalling that the cost function to be minimized (36) bears the same relationship to (41) as in the previous cases, the optimal control is once again the same as expressed by (8) and (9).

The value of the Performance Index in once again different, however, and can be shown to be [1]

$$J(u^*) = \text{trace}\,\{A \int_{0}^{\tau} e^{F\sigma}Qe^{F'\sigma}d\sigma\} + \text{trace}\,\{We^{F\tau}\bar{P}H'R^{-1}H\bar{P}e^{F'\tau}\} + \text{trace}\,\{A \int_{0}^{\tau} e^{F\sigma}\bar{P}H'R^{-1}H\bar{P}e^{F'\sigma}d\sigma\} \qquad (42)$$

In summary, the Kalman filter is used to generate, at time t, $\hat{x}(t - \tau)$. The linear predictor operates on $\hat{x}(t - \tau)$ to generate $\hat{x}(t)$. The optimal controller then operates on $\hat{x}(t)$ to generate $u^*(t)$. The total system is shown in Figure 6.

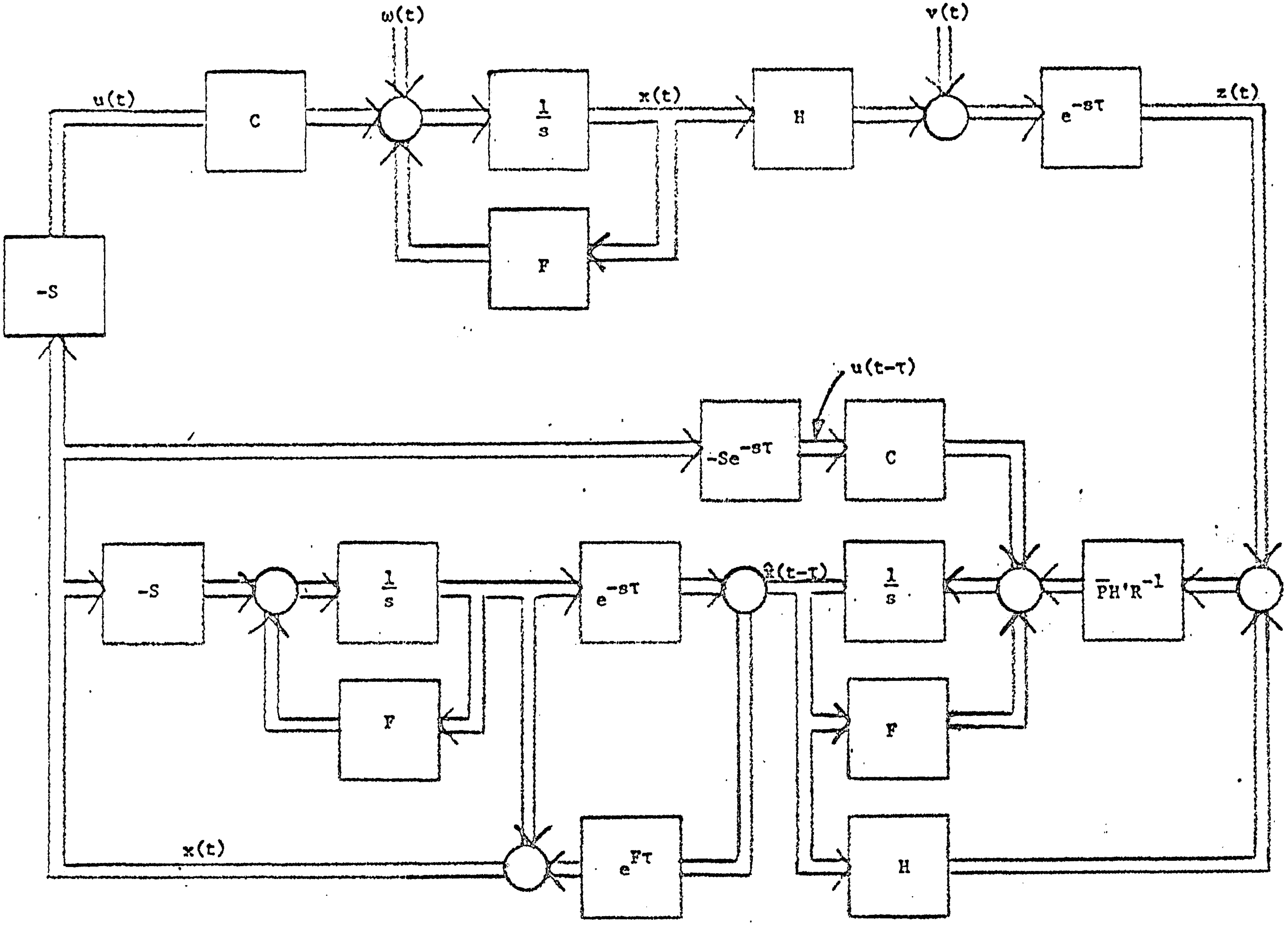


Figure 6. Optimal Controller for Plant with Measurement Noise, Plant Noise and Time Delay

APPENDIX

APPLICATION OF OPTIMAL CONTROL THEORY TO JET ENGINE PROBLEM

The results of the preceding section provide an optimal control policy for the jet engine whose transfer function is shown in block diagram in Figure 1. The plant transfer functions $G_1(s)$ and $G_2(s)$ are expressed in (1) and (2) below:

$$G_1(s) = \frac{1}{\frac{s^2}{[2(100)]^2} + s\,\frac{2(.5)}{2\pi 100} + 1} \qquad (1)$$

= BYPASS DOOR DYNAMICS

$$G_2(s) = \frac{e^{-.004s}}{\left(\frac{s}{80} + 1\right)\left(\frac{s^2}{(365)^2} + s\,\frac{2(.3)}{365} + 1\right)} \qquad (2)$$

= INLET DYNAMICS

To reduce computation time and complexity it was decided to ignore, for the present, the bypass door dynamics, $G_1(s)$, and to determine the optimal control of the plant represented by $G_2(s)$ alone. Once the control for $G_2(s)$ is known, it is then possible to either utilize the poles of $G_1(s)$ if they fall close to the poles in the desired feedback transfer function H(s), or cancel the poles of $G_1(s)$ if they occur at unwanted locations.

Thus the plant to be controlled is described by (2). The general form of the two factors of (2) is given by (3) and shown in block diagram in Figure 2. Figure 3 substitutes the appropriate numerical values as received from Lewis.

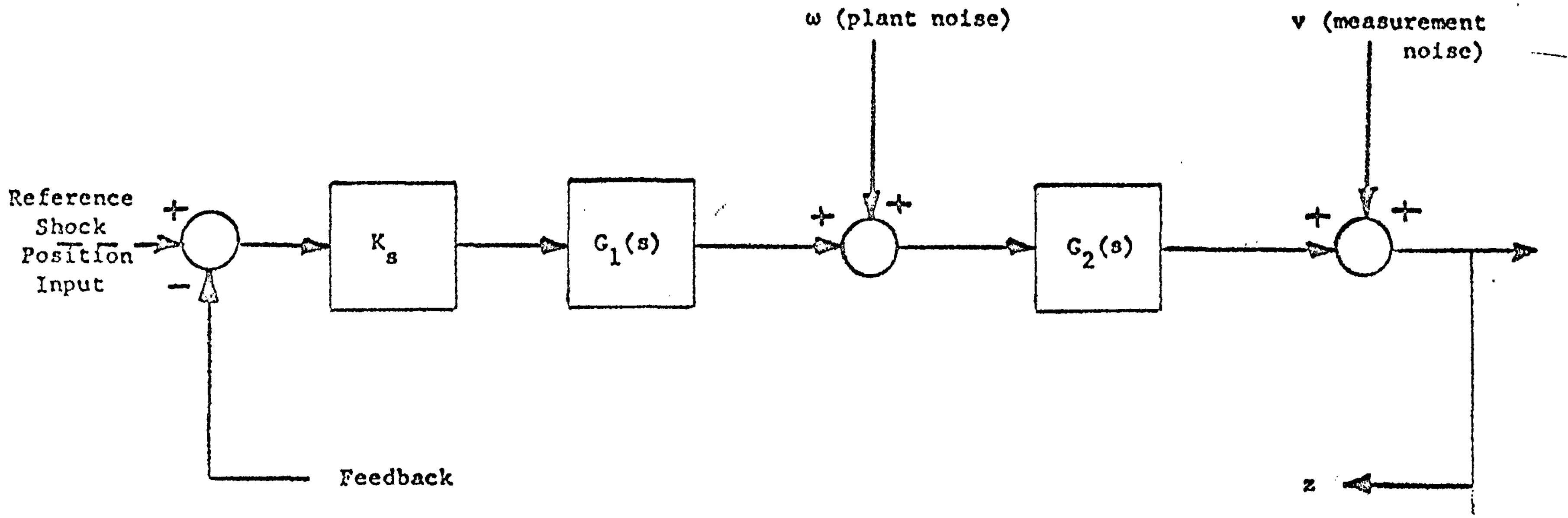
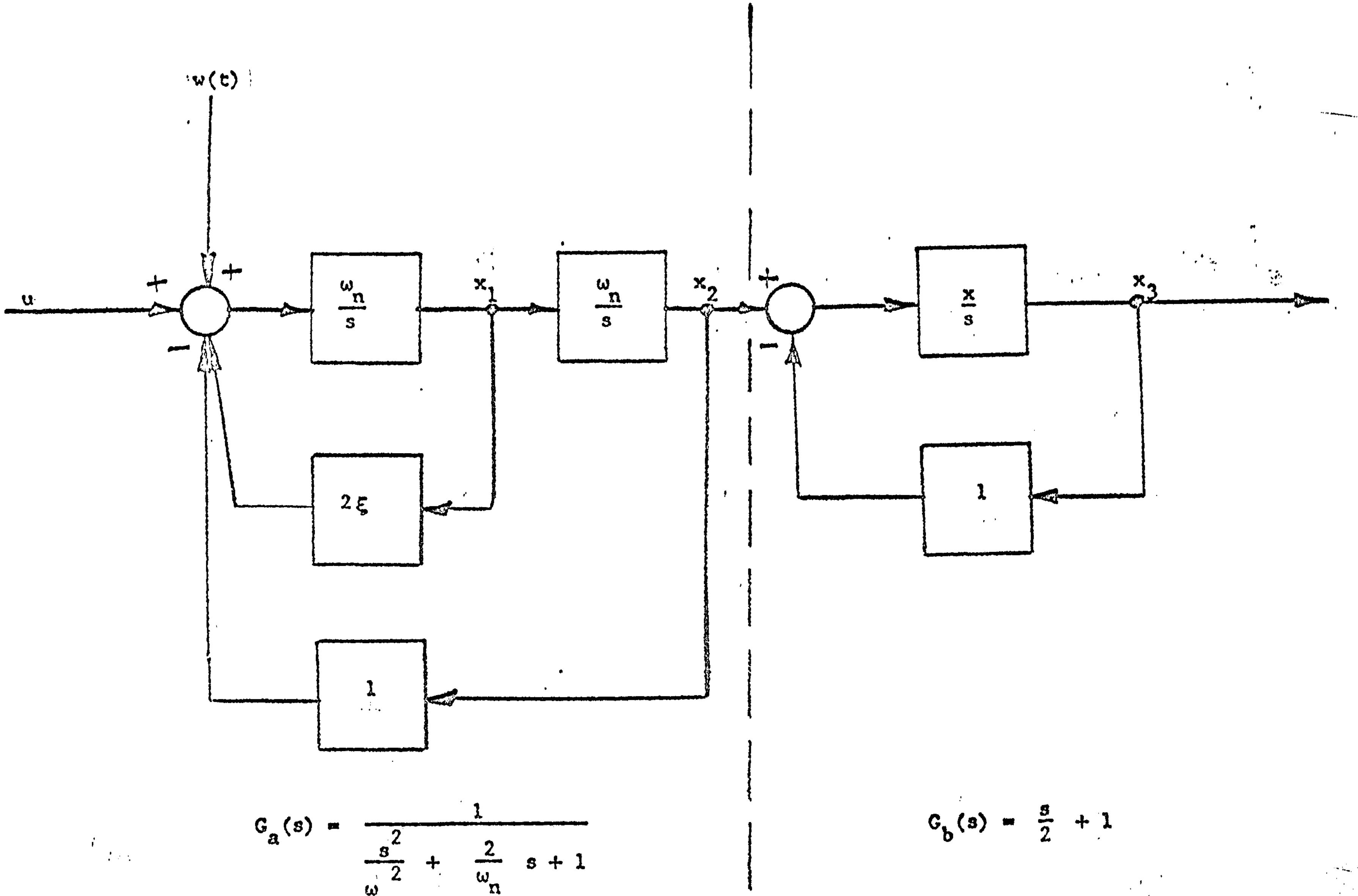


Figure 1. Inlet Block Diagram



$$G_a(s) = \frac{1}{\frac{s^2}{\omega_n^2} + \frac{2}{\omega_n} s + 1}$$

$$G_b(s) = \frac{s}{2} + 1$$

Figure 2. General Block Diagram of $G_2(s)$

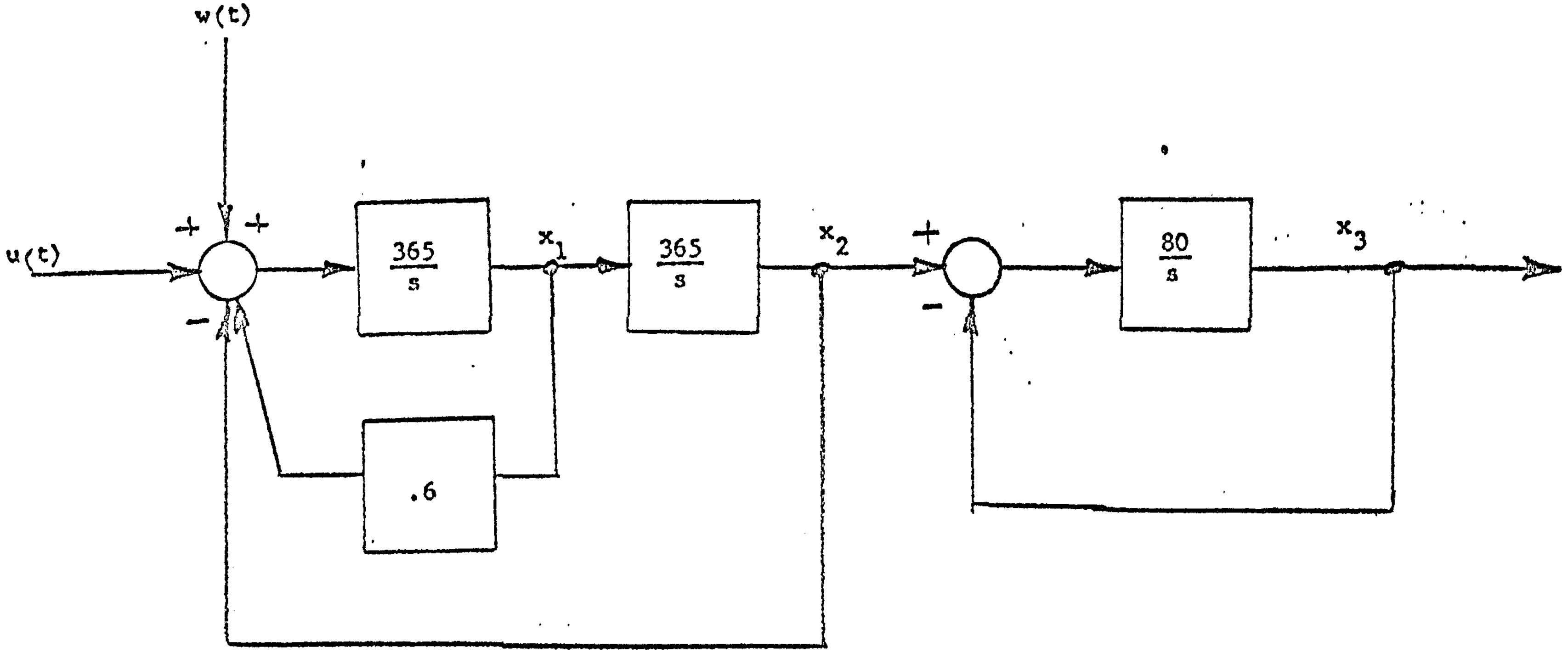


Figure 3. Numerical Block Diagram of $G_2(s)$

$$\left[\frac{1}{\frac{s}{\alpha}+1}\right]\left[\frac{1}{\frac{s^2}{\omega_n^2}+\frac{s2\xi}{\omega_n}+1}\right] \tag{3}$$

The block diagram showing the optimal control of such a plant is given in Figure 6, page 14 of this report. Note that the solution requires the determination of W(t) as defined in equation (10), page 4 and P(t) as defined in equation (32), page 10. Both (10) and (32) involve solution of integral type equations. The first attempt to solve digitally for W(t) and P(t) a one step approximation was used

$$W(t+\Delta t) = W(t) + \dot{W}(t)\Delta t \tag{4}$$

and

$$P(t+\Delta t) = P(t) + \dot{P}(t)\Delta t \tag{5}$$

Because of difficulties encountered in achieving convergence, a five-point Runge-Kutta routine was implemented. Once again great difficulty was experienced in obtaining convergence as well as excessive computation time.

A decision was then made to solve the equivalent discrete time estimation and control problems. The existence of a functional relationship between the discrete-time solutions and the continuous time solutions suggested the practicality of this approach[4].

Discrete Time Estimation

The solution to discrete time estimation is achieved by iterative solutions of (6), (7) and (8). These equations are given in Meditch[4], page 174.

$$K(k+1) = P(k+1|k)H'(k+1)[H(k+1)P(k+1|k)H'(k+1) + R(k+1)]^{-1} \tag{6}$$

$$P(k+1|k) = \Phi(k+1,k)P(k|k)\Phi(k+1,k) + \Gamma(k+1,k)Q(k)\Gamma'(k+1,k) \tag{7}$$

$$P(k+1|k+1) = [I - K(k+1)H(k+1)]P(k+1|k) \tag{8}$$

with

$$P(0|0) = E\{x(0)x'(0)\}$$

The discrete model of the plant and measurement processes is described by

$$x(k+1) = \Phi(k+1,k)x(k) + \Gamma(k+1,k)\omega(k) \tag*{( )}$$

$$z(k+1) = H(k+1)x(k) + v(k+1) \tag{10}$$

and

$$E[\omega(j)\omega'(k)] = Q(k)\delta_{jk} \tag{11}$$

A typical computational cycle proceeds as follows:

1. Given $P(k|k)$, $Q(k)$, $\Phi(k+1,k)$ and $\Gamma(k+1,k)$, $P(k+1|k)$ is computed using (7).
2. $P(k+1|k)$, $H(k+1)$ and $R(k+1)$ are then substituted into (6) to obtain $K(k+1)$.
3. $P(k+1|k)$, $K(k+1)$ and $H(k+1)$ are substituted into (8) to obtain $P(k+1|k+1)$.
4. The cycle is then repeated.

It can be shown that

$$P(t) = \lim_{\Delta t \to 0} \frac{P(t|t)}{\Delta t}$$

This fact was used by dividing by Δt the final value of $P(t|t)$ obtained by the iterative process above and substituting that result as initial conditions on $P(t)$ in equation (32), page 10. Convergence was then easily obtained for $P(t)$ in the continuous case. The resulting $P(t|t)$ for discrete time solution using time increments of 0.1 millisecond was within 2% of the continuous time solution shown in Table 1.

TABLE 1

Comparison of Discrete Estimation with Continuous Estimation

$(Q = .05, \bar{R} = .001)$

<u>Discrete Case ($\Delta t = .0001$ second) Steady State</u>

$$P(t|t) = \begin{bmatrix} 1.36944E{+}02 & 9.89223E{-}02 & -1.51184E{-}02 \\ 9.89223E{-}02 & 1.02189E{-}03 & 1.37105E{-}04 \\ -1.51184E{-}02 & 1.37105E{-}04 & 3.91911E{-}05 \end{bmatrix}$$

$$K(t+\Delta t) = \left. \begin{matrix} -1.51184E{+}01 \\ 1.37105E{-}01 \\ 3.91911E{-}02 \end{matrix} \right]$$

<u>Continuous Case - Steady State</u>

$$P(t) = \begin{bmatrix} 1.37079E{+}06 & 9.78381E{+}02 & -1.54231E{+}02 \\ 9.78381E{+}02 & 1.03177E{+}01 & 1.39888E{+}00 \\ -1.54231E{+}02 & 1.39888E{+}00 & 3.99813E{-}01 \end{bmatrix}$$

$$K(t) = \left. \begin{matrix} -1.54231E{+}05 \\ 1.39888E{+}03 \\ 3.99813E{+}01 \end{matrix} \right]$$

Discrete Time Control

A similar approach was used for solving for W(t) in the optimal control equation (10), page 4.

The discrete time equations to be solved iteratively are (9) and (10)

$$S(k) = -[\Psi'(k+1,k)W(k+1)\Psi(k+1,k) + B(k)]^{-1}x$$
$$[\Psi'(k+1,k)W(k+1)\Phi(k+1,k)] \qquad (9)$$
$$W(k) = \Phi'(k+1,k)W(k+1)\Phi(k+1,)$$
$$+ \Phi'(k+1,k)W(k+1)\Psi(k+1,k)S(k) + A(k) \qquad (10)$$

The discrete model of the plant and measurement processes now include the effect of control and are given by

$$x(k+1) = \Phi(k+1,k)x(k) + \Gamma(k+1,k)\omega(k) + \Psi(k+1,k)u(k) \qquad (11)$$
$$z(k+1) = H(k+1)x(k+1) + v(k+1) \qquad (12)$$

The performance measure J_N is quadratic of the form

$$J_N = E\{ \sum_{i=1}^{N} [x'(i)A(i)x(i) + u'(i-1)B(i-1)u(i-1)u(i-1)]\} \qquad (13)$$

Equations (9) and (10) are solved "backward" in time. That is, solution is obtained for S(N) → W(N) → S(N-1) → W(N-1) → S(N-2) ... S(N) → W(1) → S(0). A typical solution curve for a one-state problem is shown in Figure 4. It was observed that

$$\lim_{\Delta t \to 0} W(t+\Delta t)\ \Delta t = W(t)$$

The steady state values of the elements of $W(t+\Delta t)\Delta t$ were then substituted in the optimal control equation (10) in continuous time. Once again convergence was easily obtained and results are tabulated in Table 1.

In summary, the discrete time estimation and control problems were solved to avoid convergence problems encountered in solving the continuous time case. Values obtained for W(t) and P(t) were then substituted in the continuous time equations and solutions obtained.

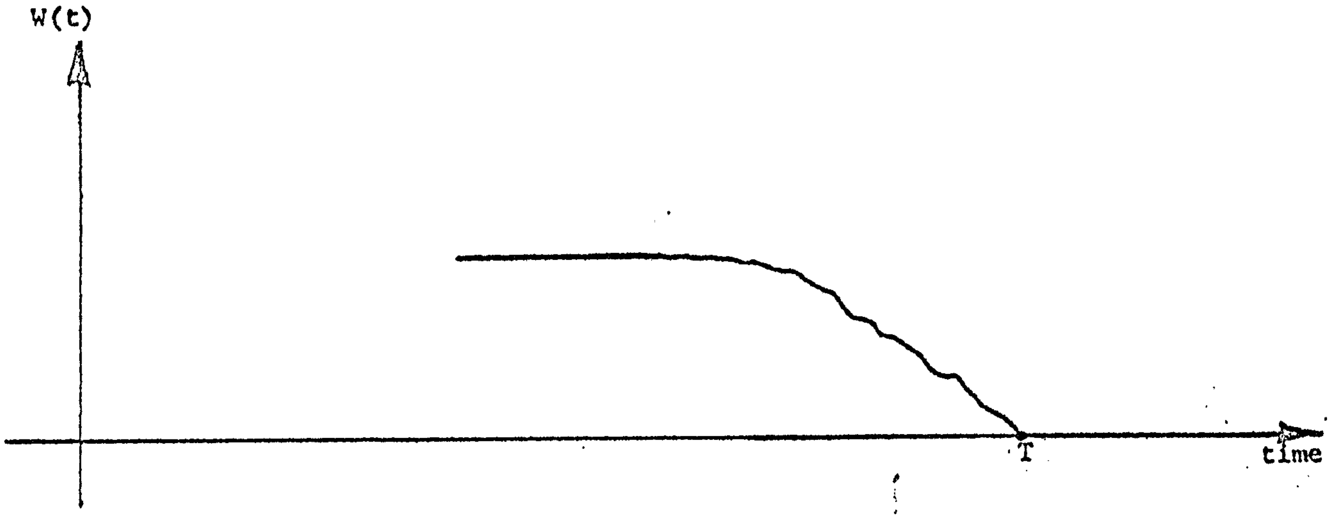


Figure 4. Typical Solution to a One-state Control Problem

Table 2

Comparison of Discrete Control and Continuous Control

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad B = [1]$$

Discrete Case ($\Delta t =$ .0001 second) Steady State

$$W_{disc} = \begin{bmatrix} 2.14491E-05 & 5.10425E-03 & 2.60397E-02 \\ 5.10425E-03 & 2.08104E+00 & 8.77318E+00 \\ 2.60397E-02 & 8.77318E+00 & 5.55050E+01 \end{bmatrix}$$

$$S_{disc} = \begin{bmatrix} -2.84938E-04 & -6.72419E-02 & -3.44291E-01 \end{bmatrix}$$

Continuous Case - Steady State

$$W_{cont} = \begin{bmatrix} 2.14420E-09 & 5.10039E-07 & 2.60386E-06 \\ 5.10039E-07 & 2.08096E-04 & 8.77395E-04 \\ 2.60386E-06 & 8.77395E-04 & 5.50037E-03 \end{bmatrix}$$

$$S_{cont} = \begin{bmatrix} -2.85178E-04 & -6.78352E-02 & -3.46312E-01 \end{bmatrix}$$

REFERENCES

1. Kleinman, D. L. "Optimal Control of Linear Systems with Time-Delay and Observation Noise," IEEE Transactions on Automatic Control, pp. 524-527, October, 1969.

2. Wonham, W. M. "Stochastic Problems in Optimal Control," Research Institute for Advanced Studies, Baltimore, Maryland, Tech. Rept. 63-14, 1963.

3. Davenport, W. and W. Root. Random Signals and Noise, New York: McGraw-Hill, 1958.

4. Meditch, J. S. Stochastic Optimal Linear Estimation and Control, New York: McGraw-Hill, 1969.

5. Wonham, W. M. "On the Separation Theorem of Stochastic Control," SIAM J. Control, Vol. 6, pp. 312-326, 1968.

UNSTART PROBLEM

R. T. Stefani

Introduction

The basic problem is to select the feedback controller for the system in Figure 1 in such a way that the following objective function is minimized

$$\lambda = \frac{\frac{1}{2\pi}\sqrt{\frac{\sigma_{\dot{y}}^{2}}{\sigma_{y}^{2}}}\; e^{-\frac{(\mathrm{TOL})^{2}}{2\sigma_{y}^{2}}}}{\frac{2}{\sqrt{2\pi}}\int_{0}^{\frac{\mathrm{TOL}}{\sigma_{y}}} e^{-\frac{x^{2}}{2}}\,dx} \qquad (1)$$

In the above, λ is the expected number of unstarts per second, y refers to the shock wave position (x_S) which is commanded to be zero, and TOL is the tolerance. This report discusses the computation of λ, the constraints on the feedback control, and the resulting design of a feedback controller. Two designs are presented: one where the feedback consists of a gain term and two real zeros and a second design using a gain term, three real zeros, and one real pole.

Calculating λ

In order to calculate λ (for the purpose of evaluating the effectiveness of any candidate feedback controller) one must obtain σ_y^2 and $\sigma_{\dot{y}}^2$, the mean square error terms. Let us consider the evaluation of σ_y^2. Hence, we wish to calculate

$$\sigma_y^2 = E\{y^2(t)\} \tag{2}$$

In order to calculate this variance, let us make use of the power spectral density which is the Fourier transform of the autocorrelation of y(t + T) and y(t).

$$S_{yy}(\omega) = \text{Fourier transform of } [E\{y(t+T)y(t)\} = R_{yy}(\tau)]$$

$$= \int_{-\infty}^{\infty} R_{yy}(t)e^{-j\omega\tau}\, d\tau \tag{3}$$

If $S_{yy}(\omega)$ is available, then σ_y^2 follows directly since

$$\sigma_y^2 = R_{yy}(0) = \frac{1}{2\pi j}\int_{-\infty}^{\infty} S_{yy}(\omega)\, dj\omega \tag{4}$$

Suppose we redraw Figure 1 into the form shown in Figure 2, thus using the superposition of the signals v and w for this linear system to obtain the response of y for zero input. Employing convolution integrals, one may obtain $S_{yy}(\omega)$ from Figure 2

$$\begin{aligned} S_{yy}(\omega) &= H_1(j\omega)H_1(-j\omega)S_{vv}(\omega) + H_2(j\omega)H_2(-j\omega)S_{ww}(\omega) \\ &\quad + H_1(j\omega)H_2(-j\omega)S_{vw}(\omega) + H_2(j\omega)H_1(-j\omega)S_{wv}(\omega) \end{aligned} \tag{5}$$

For the current problem, v and w are uncorrelated white noises, hence $S_{vw}(\omega) = S_{wv}(\omega) = 0$ and $S_{vv}(\omega)$ and $S_{ww}(\omega)$ are constants. Using this statistical knowledge, σ_y^2 can be evaluated from (4) and (5)

$$\sigma_y^2 = \frac{S_{vv}}{2\pi j}\int_{-\infty}^{\infty} H_1(j\omega)H_1(-j\omega)\, dj\omega + \frac{S_{ww}}{2\pi j}\int_{-\infty}^{\infty} H_2(j\omega)H_2(-j\omega)\, dj\omega \tag{6}$$

It immediately follows from Figure 2 and (6) that

$$\sigma_{\dot{y}}^{2} = \frac{S_{vv}}{2\pi j} \int_{-\infty}^{\infty} [j\omega H_1(+j\omega)][-j\omega H_2(-j\omega)]\, dj\omega \qquad (7)$$

$$+ \frac{S_{ww}}{2\pi j} \int_{-\infty}^{\infty} [j\omega H_2(j\omega)][-j\omega H_2(j\omega)]\, dj\omega$$

Letting $s = j\omega$, one can evaluate (7) by determining $H_1(s)$ and $H_2(s)$ from Figures 1 and 2, knowing $G_1(s)$, $G_2(s)$, and $H(s)$. One simply substitutes the correct polynomial of s into the following where the subscripts N and D refer to numerator and denominator polynomials.

$$H_1 = \frac{KK_s\ H_N\ G_{1N}\ G_{2N}}{H_D\ G_{1D}\ G_{2D} + KK_s\ H_N\ G_{1N}\ G_{2N}}$$

$$H_2 = \frac{H_D\ G_{1D}\ G_{2N}}{H_D\ G_{1D}\ G_{2D} + KK_s\ H_N\ G_{1N}\ G_{2N}} \qquad (8)$$

Constraints on the Feedback Controller

The effectiveness of any feedback controller selection is determined by evaluating λ which, in turn, requires evaluating σ_y^2 and $\sigma_{\dot{y}}^2$ using (6) and (7). In order to evaluate the required integrals, it must be true from (6) that the numerators of H_1 and H_2 are at least one order less than the corresponding denominators. It must similarly be true from (7) that the numerators of H_1 and H_2 are at least 2 orders less than the corresponding denominators. For both (6) and (7) to be calculable, the latter requirement must hold. The order of the polynomials in (8) are

Polynomial	Order
G_{1N}	0
G_{1D}	2
G_{2N}	1
G_{2D}	3
H_N	O_{H_N}
H_D	O_{H_D}

(9)

Hence H_N and H_D are, for the moment, of unknown order. If the order of the numerator of H_1 is at least 2 less than the denominator, then from (8) and (9)

$$[(O_{H_N} + 1) - (O_{H_D} + 5)] \leq -2$$

That is, upon simplifying

$$O_{H_N} - O_{H_D} \leq 2 \qquad (10)$$

Similarly, if the order of the numerator of H_2 is at least 2 less than the denominator, from (8) and (9) we have

$$[(O_{H_D} + 3) - (O_{H_D} + 5)] \leq -2 \qquad (11)$$

which obviously holds for any order of H_D. We conclude that any candidate feedback controller can have a zero over poles excess of no more than 2. Furthermore, the poles of (8) must all be in the left half plane. Using root locus considerations we note that the number of open loop poles is $5 + O_{H_D}$ while the number of open loop zeros is $1 + O_{H_N}$. Hence, the pole over zero excess is $4 + (O_{H_D} - O_{H_N})$.

If we wish to insure that the closed loop poles are all in the left half plane for any gain choice, we can do so by selecting the maximum zero over pole excess for the feedback controller thus providing a net open loop pole over zero excess of 2. The result is a 90° asymptote whose real crossing may be kept in the left half plane by properly choosing the feedback controller poles and zeros. That is

$$0 \geq \text{real axis crossing} = \frac{1}{2}\{\Sigma \text{ real parts of poles of } G_{1D}, G_{2D} - \text{zero of } G_{2N} + \Sigma \text{ zeros of } H_D - \Sigma \text{ zeros of } H_N\} \qquad (12)$$

Figure 3 contains the selected transfer functions for G_1 and G_2. The result of using the transfer functions of Figure 3 with (12) is the inequality

$$\Sigma \text{ zeros of } H_D - \Sigma \text{ zeros of } H_N \leq 717 \text{ rad./sec.} \qquad (13)$$

From the above discussion

$$O_{H_N} - O_{H_D} = 2 \qquad (14)$$

Finally, to insure that no locus may result in a right half plane pole

$$\begin{aligned} &\text{zeros of } H_D \leq 0 \\ &\text{zeros of } H_N \leq 0 \end{aligned} \qquad (15)$$

In summary, (13)-(15) provide constraints on the feedback controller such that the resulting system is stable and the transfer functions H_1 and H_2 have numerators of order 2 less than the denominators all of which are necessary such that the objective function λ may be calculated.

Obtaining the Optimal Feedback Controller

In order to obtain the optimal feedback controller, two problems must be considered. The first problem is calculating the objective function λ while the second problem is varying the feedback parameters such that the optimal (minimum) value of λ is achieved.

Calculating λ is discussed in the last two sections. One must solve (6) and (7) where H_1 and H_2 are defined in (8). A subroutine INTSQ is used to evaluate integrals of the form

$$I = \frac{1}{2\pi j}\int_{-\infty}^{\infty} F(j\omega)F(-j\omega)\,dj\omega \qquad (16)$$

INTSQ is discussed in the appendix. If conditions (13)-(15) are met, then $F(j\omega)$ in (16) satisfies the requirements of INTSQ.

The problem of varying the parameters to optimize λ is solved by using a subroutine designed for one function of two variables. Since the problem at hand treats more than two variables, at each step in the process all but two variables must be fixed while the remaining two are varied. The subroutine is PLOT3D, so named because a three dimensional plot is obtained, that is, a plot of the function versus two independent variables. To utilize PLOT3D, 100 values of each of the two variables are selected and the objective function is evaluated at all 10,000 combinations. The subroutine quantizes the data into 26 levels of ascending magnitude. PLOTØD then prints out a 100 by 100 array of letters A-Z. The horizontal and vertical axes represent the independent variables while the letters represent the magnitude of the objective function. In essence one has a contour plot of the function over the selected parameter range. One can then select the optimal value if it is interior to the plot or make a judgment from the contours as to what new parameter range is required to obtain the optimal value.

In Figure 4 a general block diagram of the computer program is shown. The program must be given values for all but two of the feedback parameters. Then, 100 values are selected for each of the other two parameters. UNSTART is called to evaluate λ. The subroutine UNSTART, in turn, calls INTSQ to evaluate (6) and (7), evaluates λ from (1), and returns λ to the main program. When all 10,000 values have been obtained, PLOT3D is called to furnish the 100 by 100 contour plot of the function.

Once the optimal feedback parameters are available, one can determine the closed loop poles by using a root solving scheme.

Plan of Attack for the UNSTART Problem

The following plan of attack was used for the unstart problem. A feedback controller consisting of two zeros and a gain was selected first.

$$H(s) = K(s + a)(s + b) \tag{17}$$

The parameter b was fixed while a and K were varied to obtain an optimal K. Then K was fixed at the resulting optimal and a and b were varied to obtain optimal choices. Next a and b were fixed at the resulting optimal values and a new optimal K was chosen. Finally, the new optimal K was used to obtain a new optimal a,b pair. Hence two cycles of computation were made. The closed loop poles were then obtained for the resulting values of K, a, and b.

A second feedback controller was considered

$$H(s) = \frac{K(s + a)(s + b)(s + c)}{s + d} \tag{18}$$

The previously selected values of K, a, and b were used. Parameters c and d were varied and an optimal pair was found. The resulting closed loop poles were obtained. In regard to (17) and (18) it is necessary to satisfy (13)-(15) so that all integrals calculated by subroutine INTSQ are valid and so that the resulting system is stable. In terms of the parameters a, b, c, and d, one must have, to satisfy (13)-(15)

$$\begin{aligned} a, b, c, d &\geq 0 \\ a + b + c - d &\leq 717 \text{ rad./sec.} \end{aligned} \tag{19}$$

Optimal Feedback Controller $H(s) = K(s + a)(s + b)$

The procedure discussed above was used to evaluate optimal values for K, a, and b. With b arbitrarily fixed at 23, a plot was made of the contours of λ for various values of K and a as shown in Figure 5 for exponential variations of K and a. There are two interesting points regarding Figure 5. Note first that the optimal gain selection is constant over a large range of a. As a result $K = K_1 = 8.84 \times 10^{-6}$ was selected. The second interesting point regards the ridge. This ridge is, in effect, a stability contour. On one side one calculates invalid values of λ whereas on the other side (for a stable system) one has valid λ values. If one ignored the necessity of maintaining stability, one might be tempted to seek the invalid minimum. Furthermore, the stability contour (valid for various gain selections) is less restrictive of allowable values of a than is the restriction imposed by (13)-(15) which requires

$$\begin{aligned} &a, b \geq 0 \\ &a + b \leq 717 \text{ rad./sec.} \end{aligned} \tag{20}$$

This occurs since (20) is true for all gain values whereas Figure 5 treats specific gain values.

With $K = 8.84 \times 10^{-6}$ contours of λ versus a and b were obtained as in Figure 6. Note again the presence of a ridge which is a stability contour related to the above gain selection and which is less restrictive than (20). The optimal pair a, b was

$$a = b = 235 \text{ rad./sec.}$$

With $a = b = 235$ it was found that the optimal value of K was 8.9×10^{-6}. With K so chosen Figure 6 was repeated with no change in the optimal pair a, b. Hence the optimal parameters are

$$\begin{aligned} K &= 8.9 \times 10^{-6} \\ a = b &= 235 \text{ rad./sec.} \end{aligned} \tag{21}$$

The optimal objective function and mean square errors were

$$\begin{aligned} \lambda &= 73.76 \text{ unstarts/sec.} \\ \sigma_y^{\,2} &= 4.845 \text{ (in)}^2 \\ \sigma_{\dot{y}}^{\,2} &= 2.092 \times 10^6 \text{ (in.sec.)}^2 \end{aligned} \tag{22}$$

The closed loop poles were (see the root locus of Figure 7).

$$\begin{aligned} &-107 \\ &-239 \pm j223 \\ &-171 \pm j1001 \end{aligned} \tag{23}$$

Assuming that we have

a normal distribution for y and $\dot{y}$ and using a ± 2σ range we can say with 98°/o certainty that the shock wave position is between ±4.4 inches with a time rate of change between ±2890 inches/second. This oscillation is probably due to the lightly damped pole in (23). It is not obvious that a tight bound on shock wave position at the expense of a high value of $\sigma_{\dot{y}}^2$ is as detrimental as is implied by λ since λ is directly proportional to $\sigma_{\dot{y}}$.

Optimal Feedback Controller $H(s) = \frac{K(s + a)(s + b)(s + c)}{(s + d)}$

The values of K, a, and b from (21) were used. Parameters c and d were varied and the contour plot of Figure 8 was obtained. Once again the ridge separating stability regions is less selective than a plot of (19) which, for a and b chosen above, requires

$$\begin{aligned} c, d &\geq 0 \\ c - d &\leq 247 \text{ rad./sec.} \end{aligned} \tag{24}$$

The optimal choice of c and d were

$$\begin{aligned} c &= 951 \text{ rad./sec.} \\ d &= 1343 \text{ rad./sec.} \end{aligned} \tag{25}$$

The resulting optimal objective function and mean square errors are

$$\begin{aligned} \lambda &= 70.46 \text{ (unstarts/second)} \\ \sigma_y^2 &= 6.02 \text{ (in)}^2 \\ \sigma_{\dot{y}}^2 &= 1.59 \times 10^6 \text{ (in./sec.)} \end{aligned} \tag{26}$$

The closed loop poles were (see the root locus of Figure 9)

$$\begin{array}{l} -101 \\ -223 \pm j245 \\ -243 \pm j937 \\ -1238 \end{array} \qquad (27)$$

Although λ was reduced somewhat from (22) by reducing $\sigma_{\dot{y}}^2$, it is not clear that this design is better since the bound on σ_y^2 is now looser.

Summary

Optimal feedback controllers were obtained with the resulting parameters, minima, and closed loop poles contained in (21)-(23) and (25)-(27). Some consideration should be given to the actual importance of controlling $\dot{y}$ since y can be held to within reasonable bounds. It may be of interest to design the feedback controllers to minimize σ_y^2 as defined in (6) rather than λ which is defined in (1) and is directly proportional to $\sigma_{\dot{y}}$. It is concluded also that contour plots of functions such as λ (and also σ_y^2 or $\sigma_{\dot{y}}^2$) contain ridges which are, in effect, stability boundaries, hence care must be taken as to the proper direction in which to search for minima.

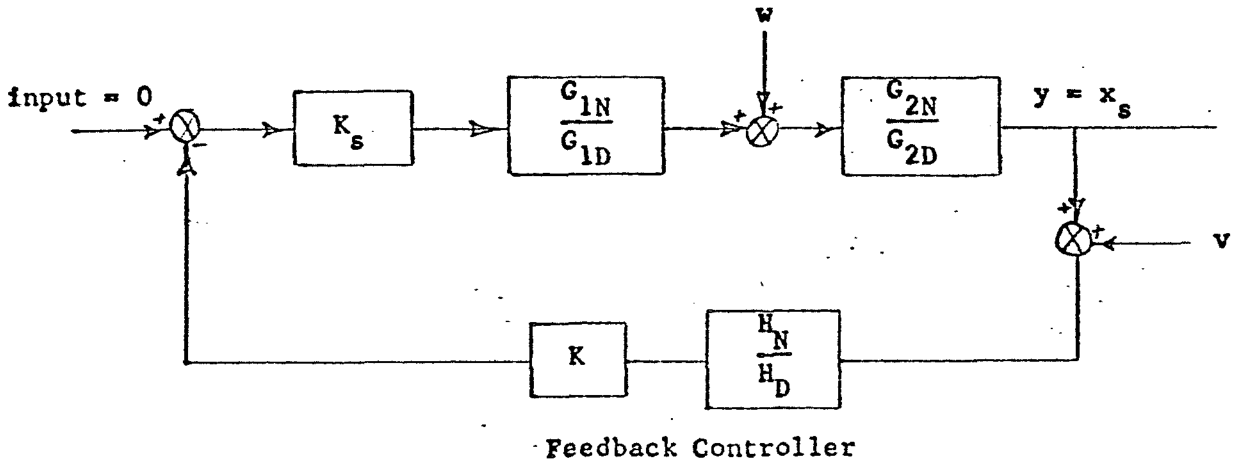


Figure 1. The Basic Control Scheme

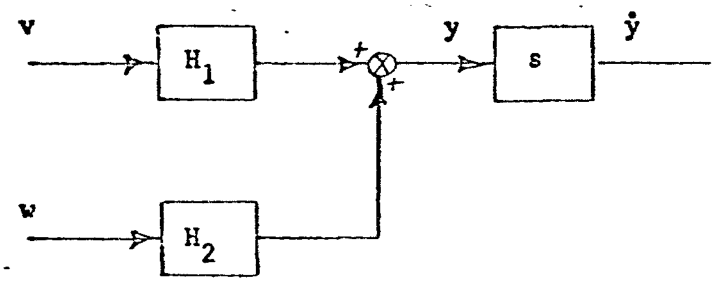


Figure 2. Figure 1 Redrawn to Obtain σ_y^2 and $\sigma_{\dot{y}}^2$

$$G_1(s) = \frac{G_{IN}}{G_{ID}} = \frac{(2\pi 100)^2}{S^2 + (2\pi 100)\ S + (2\pi 100)^2}$$

(lb/sec./volt)

$$= \frac{(2\pi 100)^2}{(S + 314.2 + j544.1)\ (S + 314.2 - j544.1)}$$

$$G_2(s) = \frac{G_{2N}}{G_{2D}} = \frac{\frac{(2.9)\ (80)(365^2)}{210} \times (s + 210)}{(s+80)\ (s^2 + (.6)\ (365)\ s + 365^2)}$$

$$= \frac{\frac{(2.9)\ (80)\ (365^2)}{210} \times (s + 210)}{(s+80)\ (s+109.4 + j\ 348.1)\ (s + 109.4 - j348.1)}$$

(in/lb/sec.)

Figure 3 - Transfer Functions For The Unstart Problem

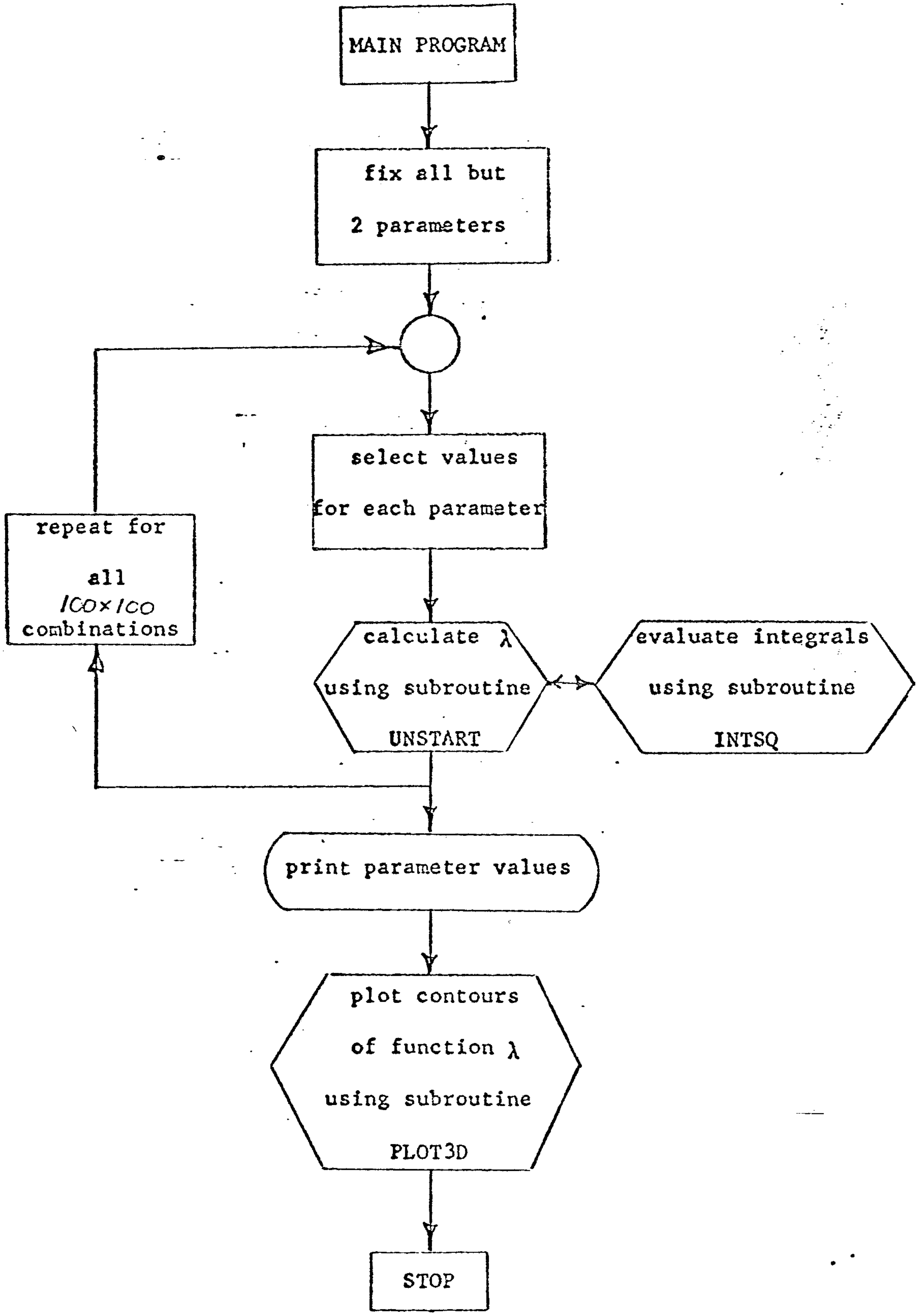

MAIN PROGRAM
fix all but 2 parameters
select values for each parameter
repeat for all 100×100 combinations
calculate λ using subroutine UNSTART
evaluate integrals using subroutine INTSQ
print parameter values
plot contours of function λ using subroutine PLOT3D
STOP

Figure 4. Computer Program Flow Diagram

λ min = 82 (sec^{-1})

λ_1= 93 (sec^{-1})

λ_2 = 120 (sec^{-1})

b = 235 (rad/sec.)

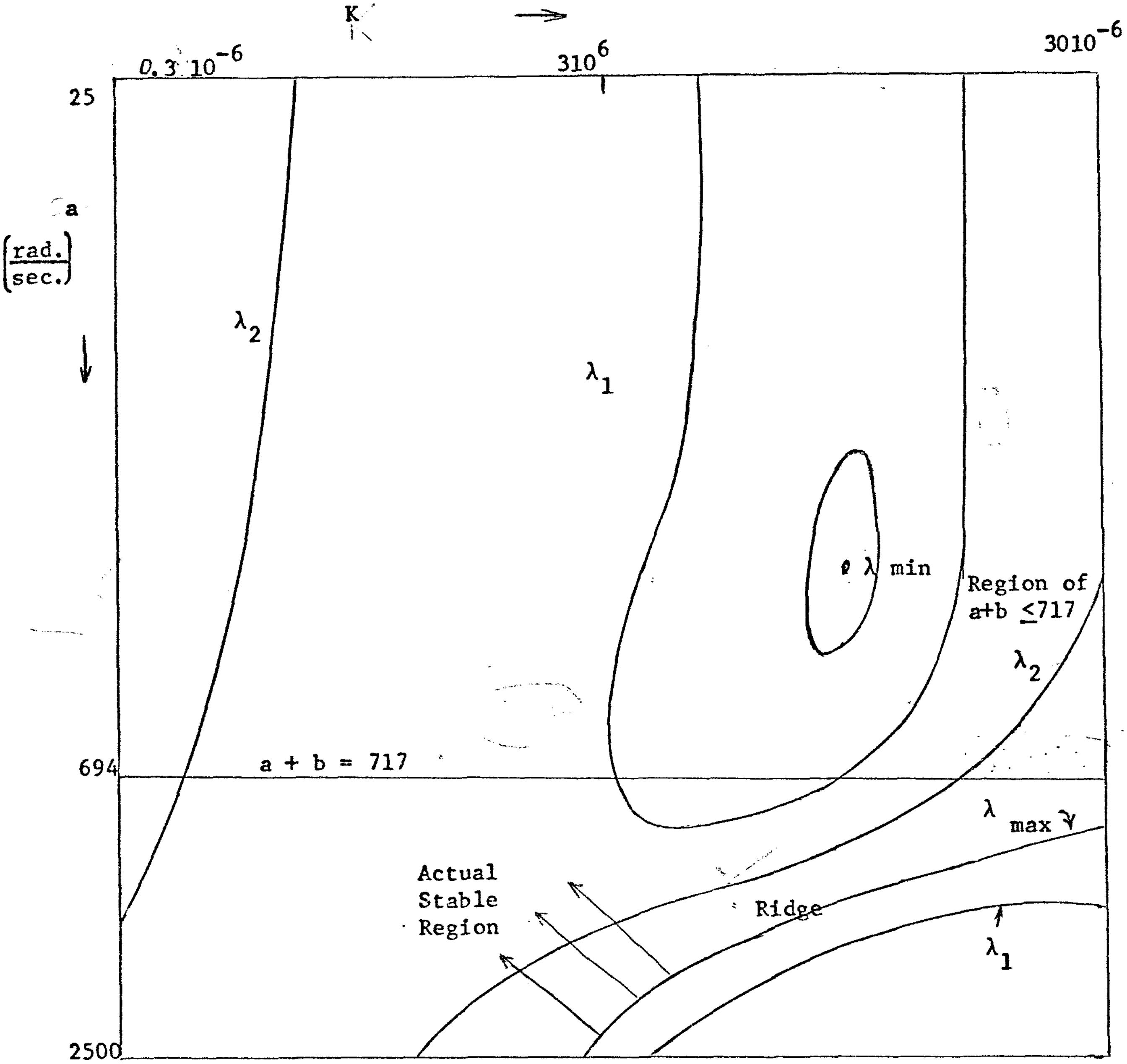


Figure 5. Contours of λ Versus K and a for Fixed b

λ min = 74 (sec^{-1})

λ_1 = 75 (sec^{-1})

λ_2 = 77 (sec^{-1})

K = 8.9 10^{-6}

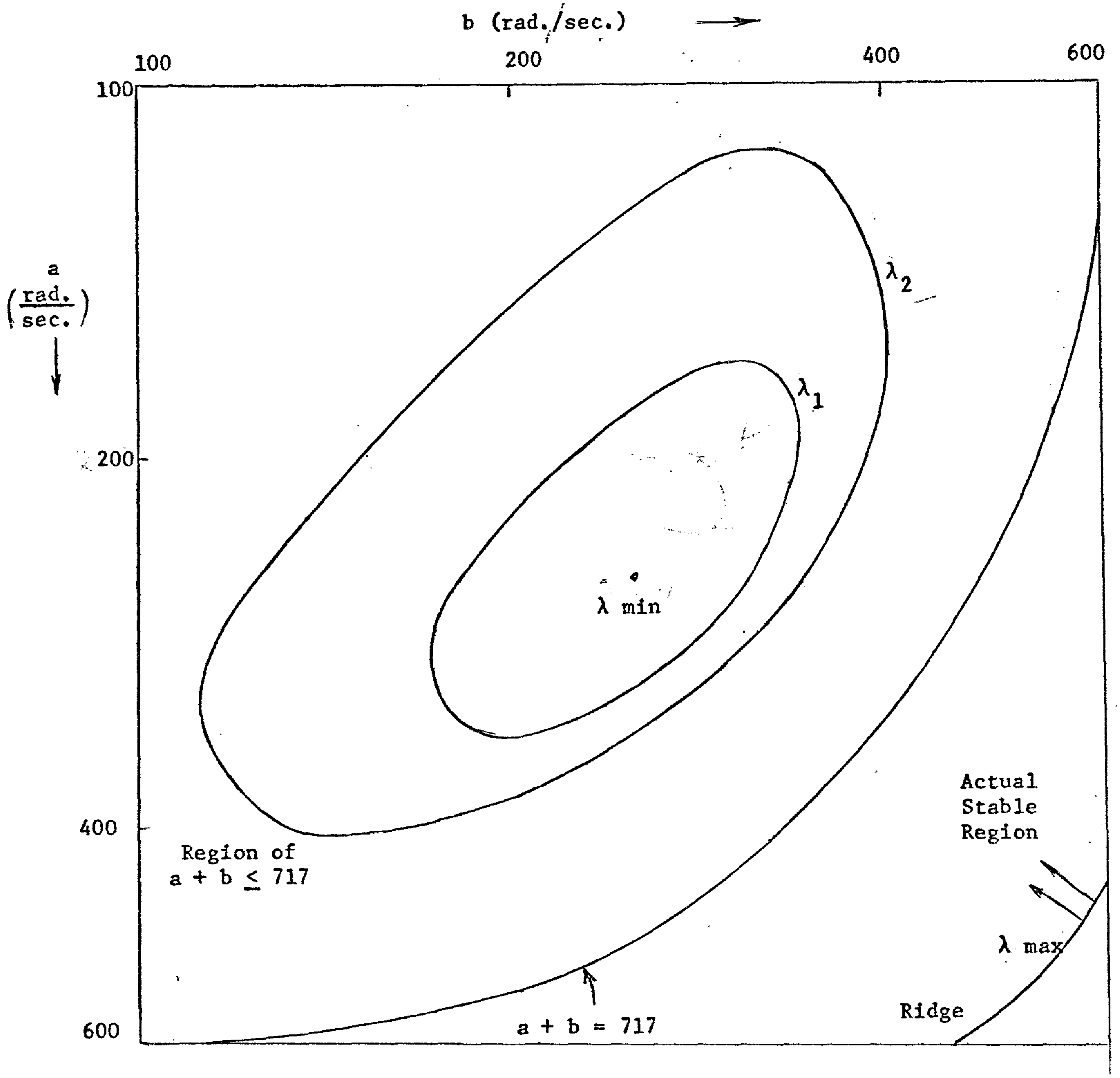


Figure 6. Contours of λ Versus a and b for Fixed K

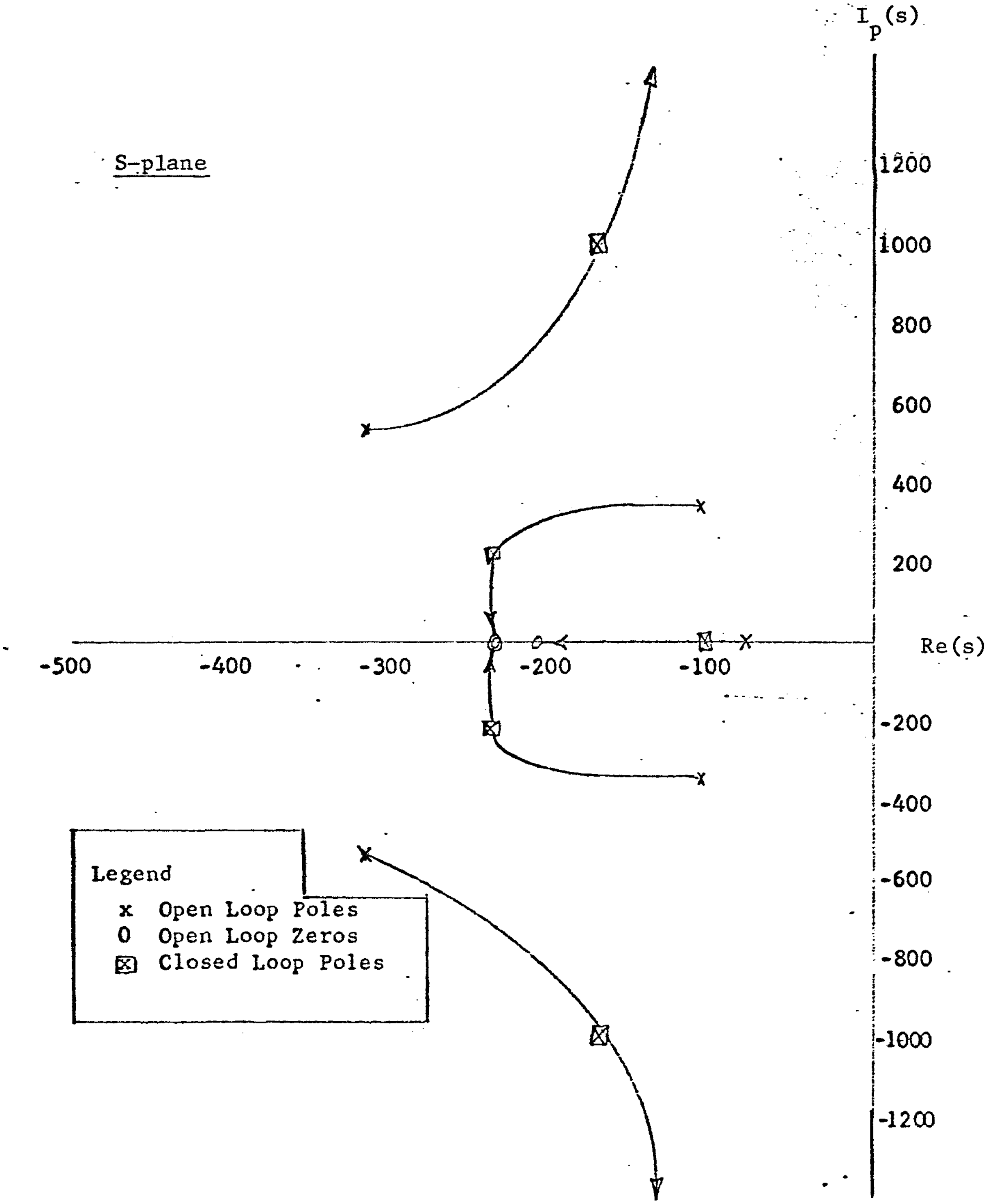


Figure 7. Root Locus H(s) = K(s+a)(s+b)

$\lambda_{min} = 58\ (\text{sec}^{-1})$

$\lambda_1 = 34\ (\text{sec}^{-1})$

$\lambda_2 = 71.5\ (\text{sec}^{-1})$

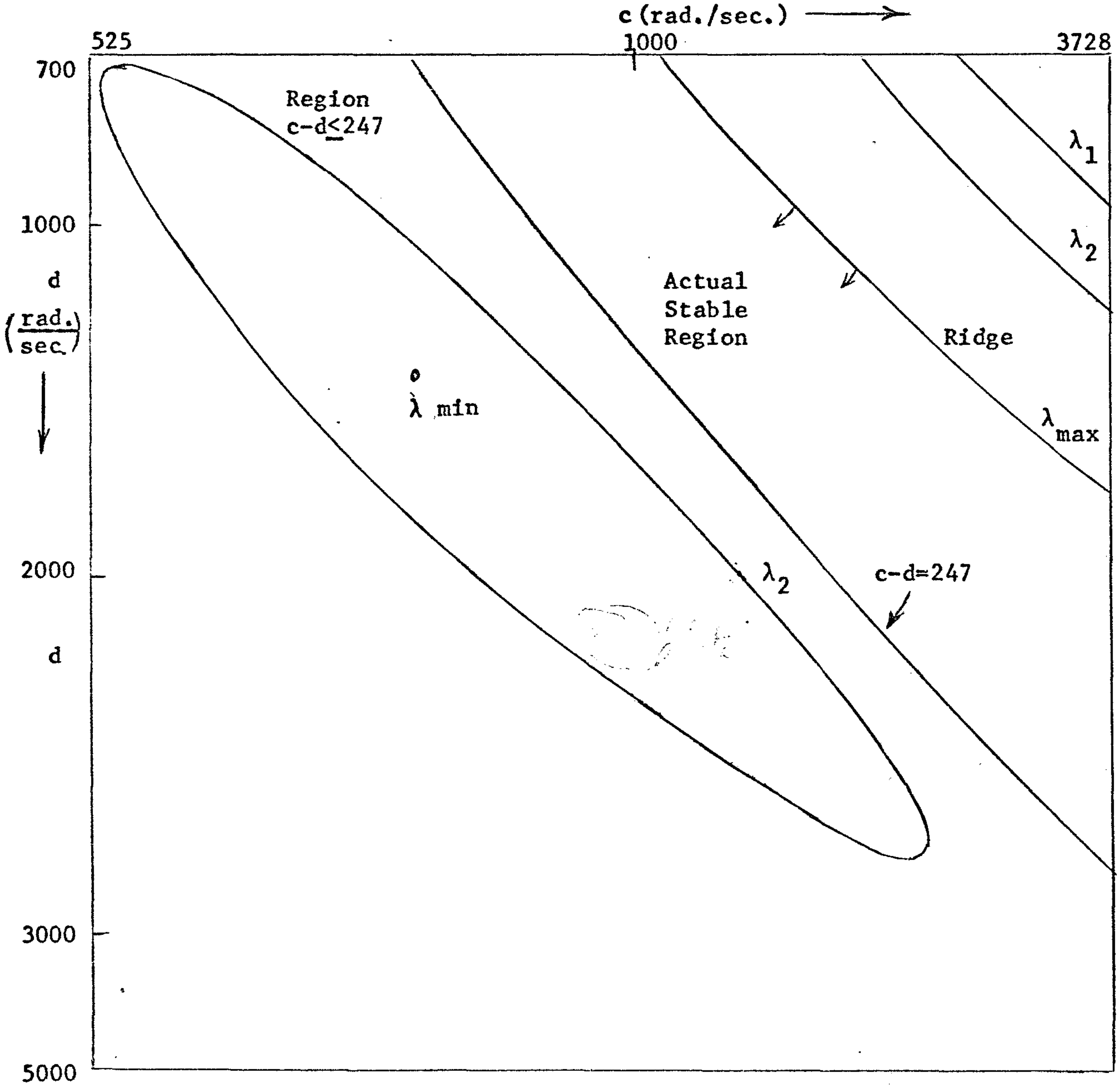


Figure 8. Contours of λ Versus c and d for Fixed K, a, and b

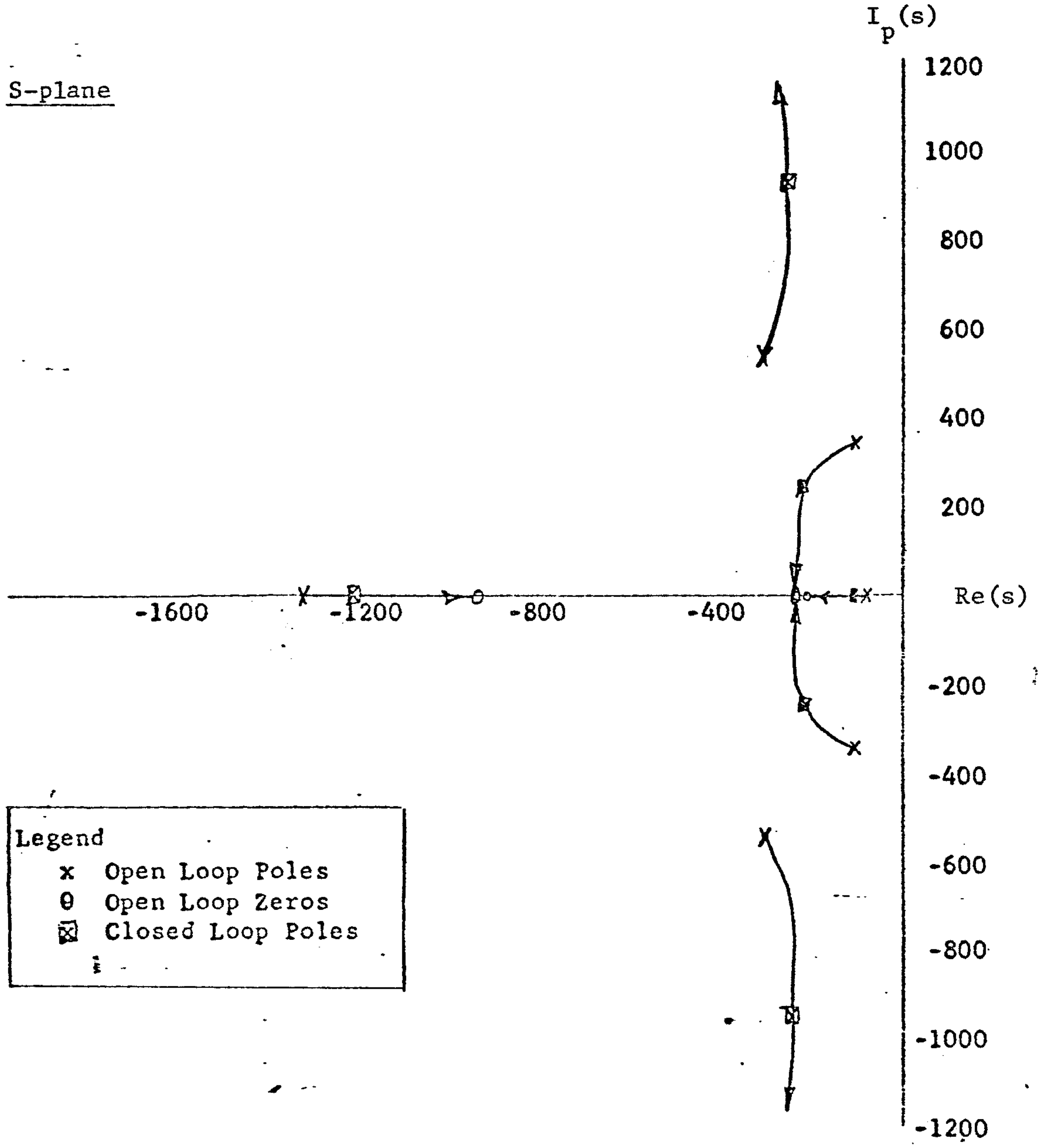


Figure 9. Root Locus for $H(s) = \frac{K(s+a)(s+b)(s+c)}{s+d}$

Appendix

A closed form solution for the integral

$$I = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} \frac{g(p)}{h(p)h(-p)}\, dp = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} F(p)F(-p)\,dp \qquad (A1)$$

is presented in the paper by F. H. Effertz.[1] The solution takes the form of Equation 4 in his paper. A better algorithm for the computation of Equation 4 in Effertz is presented in the correspondence by Pazdera[2] (Equation 7'). A modified form of Pazdera's algorithm has been coded in FORTRAN.

The use of Equation 4 in Effertz can be best illustrated by an example. Suppose we wish to evaluate the following

$$I = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} \frac{s + c}{s^2 + (a+b)s + ab} \; \frac{c - s}{s^2 - (a+b)s + ab}\, ds \qquad (A2)$$

It can be shown from residue theory that the correct answer is

$$\begin{aligned} I = {} & \text{Residue of } F(s)F(-s) \text{ at } s = -a \\ & + \text{Residue of } F(s)F(-s) \text{ at } s = -b \end{aligned} \qquad (A3)$$

where $F(s)F(-s)$ indicates the function to be integrated over $s = j\omega$. Then the answer is

$$I = \frac{c^2 + ab}{2\, ab(a + b)} \qquad (A4)$$

To use Equation 4 from Effertz we must make the following associations, using the complex frequency variable S in place of p

$$g(p) = (s + c)(c - s) = c^2 - s^2$$

$$\begin{aligned} \text{then} \quad & p = s \\ & n = 2 \\ & b_0 = -1 \\ & b_1 = c^2 \end{aligned}$$

$$h(p) = s^2 + (a + b)s + ab \tag{A5}$$

$$\begin{aligned} \text{then} \quad & p = s \\ & N = 2 \\ & a_0 = 1 \\ & a_1 = a + b \\ & a_2 = ab \end{aligned}$$

The integral, for $n = 2$, can be expressed as

$$I = \frac{(-1)^3}{2a_0} \frac{\begin{vmatrix} b_0 & b_1 \\ a_0 & a_2 \end{vmatrix}}{\begin{vmatrix} a_1 & 0 \\ a_0 & a_2 \end{vmatrix}} = \frac{-1}{2a_0} \frac{b_0a_2 - a_0b_1}{a_1 a_2} \tag{A6}$$

Substituting the correct values of a_0, a_1, etc., and cancelling minus signs for this problem we obtain

$$I = \frac{c^2 + ab}{2\ ab(a + b)} \tag{A7}$$

What is most interesting is that the result depends only on the coefficients of the known function being integrated and not on the poles of the function as one would suspect from residue theory.

It has been mentioned that the algorithm from Pazdera (Equation 7') has been programmed. Some changes in the nomenclature of Equation 7' were necessary to facilitate coding. Note that the practice in Effertz and Pazdera is to make the highest numbered coefficient correspond to the lowest power of the variable p. Note, also, that subscripts such as a_0 are in evidence. To facilitate coding, the lowest numbered coefficient was subscripted in the form A(1) and this coefficient was associated with the lowest power of the variable p, in this case p^0. In general, the i^{th} coefficient A(I) is associated with the $(i - 1)^{st}$ power of p, p^{I-1}. (In FORTRAN the expression A(0) is not allowed). Note from Equation A1 above what must be done to operate on the function F(s)F(-s). If F(s) is of the form

$$F(s) = \frac{c(s)}{A(s)} \tag{A8}$$

then the algorithms suggested in both Effertz and Pazdera require the use of

$$\begin{aligned} g(s) &= c(s)c(-s) \\ h(s) &= A(s) \end{aligned} \tag{A9}$$

It is more desirable to input C(S) and A(S) rather than C(S)C(-S) and A(S). Consequently one uses the INTSQ subroutine by coding the coefficients of C(S) and A(S) with ascending subscripts corresponding to ascending powers of S. The subroutine provides the operation C(S)C(-S). Furthermore, the subroutine checks to see if the lowest coefficient of C(S) or A(S) is zero so that factors of S may be either considered or cancelled. Two basic requirements must be met by A(S) and C(S). First, the roots of A(S) and C(S) must at least have $\geqslant 0$ real parts. (Hurwitz polynomial requirement). Secondly, the highest power of C(S) must be at least one less than the highest power of A(S) for convergence to be assured. If A(S) is Nth order, one inputs the N+1 coefficients of A(S) and the N coefficients of C(S).

The use of this subroutine INTSQ may be best illustrated by an example. Suppose we desired to evaluate the integral

$$I = \int_{-j\infty}^{j\infty} \frac{s + 4}{4s^3 + 3s^2 + 2s + 1} \quad \frac{4 - s}{-4s^3 + 3s^2 - 2s + 1} \, dS \qquad \text{(A10)}$$

The polynomials C(S) and A(S) as defined in Equation A8 are input in the program as

$$\begin{array}{ll} C(1) = 4. & A(1) = 1. \\ C(2) = 1. & A(2) = 2. \\ C(3) = 0 & A(3) = 3. \\ & A(4) = 4. \end{array} \qquad \text{(A11)}$$

The program then utilizes the modified Pazdera algorithm and prints out the message

```
THE VALUE OF THE INTEGRAL IS 12.250
```

To check this answer, one can use Effertz Equation 4 for a third order case (n = 3).

$$I = \frac{(-1)^4}{2a_0} \; \frac{\begin{vmatrix} b_0 & b_1 & b_2 \\ a_0 & a_2 & 0 \\ 0 & a_1 & a_3 \end{vmatrix}}{\begin{vmatrix} a_1 & a_3 & 0 \\ a_0 & a_2 & 0 \\ 0 & a_2 & a_3 \end{vmatrix}} \quad \text{(A12)}$$

$$= \frac{1}{2a_0} \; \frac{b_0\, a_2\, a_3 + b_2\, a_0\, a_1 - a_0\, a_3\, b_1}{a_1\, a_2\, a_3 - a_0\, a_3^2}$$

For our example problem, the following associations follow

$$\begin{array}{ll} b_0 = 0 & a_0 = 4 \\ b_1 = -1 & a_1 = 3 \\ b_2 = 16 & a_2 = 2 \\ & a_3 = 1 \end{array} \quad \text{(A13)}$$

Substituting the values of Equation A13 into Equation A12 verifies the computers solution I = 12.25.

References

1. F. H. Effertz. "On Two Coupled Matrix Algorithms for the Evaluation of RMS Error Criterion of Linear Systems," Proceedings of the IEEE, June 1966, pp. 879-880.

2. J. S. Pazdera. Correspondence, Proceedings of the IEEE, November 1966, pp. 1628-1629.

UNBIASED STRUCTURAL PARAMETER ESTIMATION

R. T. Stefani

Goal of This Study and Plan of Attack

The goal of this study is to utilize a weighted least squares objective function and formulate an estimation algorithm which is unbiased when applied to structural parameter estimation (i.e., the estimation of parameters using quantities which are observed with uncertainty while a relationship exists between the unobservables). In this case conventional weighted least squares techniques lead to biased estimates. The estimation algorithm should be applicable even when noise statistics are unknown, requiring some method of estimating the statistics.

The following plan of attack is suggested. A search will be made of related literature. The theory of stochastic processes and random variables will be applied to analyzing the convergence properties and mean square error for candidate algorithms. Linear system theory will be used to simulate (digitally) a linear system on which to compare the estimation algorithms. Conventional techniques will also be considered (e.g., conventional weighted least squares methods and the instrumental variable approach). Additional applications of the new technique will be sought, hopefully in fields which have not previously been considered.

Structural Parameter Estimation

Structural parameter estimation is best introduced by means of a simple example. Suppose there exists an exact linear relationship between quantities Y_e and X_e, that is, in matrix form

$$Y_e = X_e h \tag{1}$$

Suppose we have measurements Y_s and X_s of Y_e and X_e respectively. Further, suppose we wish to estimate h such that we also minimize

$$J = (Y_s - X_s \hat{h})^T M (Y_s - X_s \hat{h}) \tag{2}$$

Minimizing J is a weighted least squares minimization problem. By selecting $\hat{h}$ such that $\frac{\partial J}{\partial \hat{h}} = 0$, we have, assuming that M is symmetrical

$$\hat{h} = (X_s^T M X_s)^{-1} X_s^T M Y_s \tag{3}$$

Let us assume that

$$Y_s = Y_e + V = X_e h + V \tag{4}$$

where V is a noise term with zero mean and a covariance matrix R. The above problem may be considered a conventional weighted least squares problem if $X_s = X_e$. In that case, the expected value of $\hat{h}$ is h as can be seen by substituting (4) into (3) and taking the expected value with $X_s = X_e$.

However, a more general and more practical problem arises when X_e is known with uncertainty.

$$X_s = X_e + N \tag{5}$$

where N is a noise term with zero mean and a covariance matrix S.

The problem becomes one of structural parameter estimation[1] in that a structural relationship exists between the unobservables Y_e and X_e both of which are known with uncertainty. For the structural parameter estimation case, the expected value of $\hat{h}$ becomes

$$E(\hat{h}) = (X_e^T M X_e + T)^{-1} X_e^T M X_e h$$
$$T = E\{N^T M N\} \qquad (6)$$

Thus, a biased estimate occurs due to the presence of T.

Previous Approach to Bias Removal

An approach to removing the bias in (6) was discussed in the March 1970 progress report. This technique consists of changing the estimation algorithm of (3) by subtracting T (defined in (6)) as follows

$$\hat{h} = [X_s^T M X_s - T]^{-1} X_s^T M Y_s \qquad (7)$$

Upon taking the expected value of $\hat{h}$ in (7) using (4) and (5) one finds that one has an unbiased estimate of h. However, the problem immediately arises as to the selection of M (assuming one is free to choose M). Furthermore, (7) does not minimize J as required by (3).

The problem of selecting M to minimize the variance of the estimation error was discussed in the March 1970 progress report. Recent work has indicated that further efforts to that end will be fruitless. The existence of a best linear unbiased estimate for h requires, as per the Gauss-Markov theorem,[2] that

$$E(Y_s) = X_s h \qquad (8)$$

In the case at hand

$$E(Y_s) = X_e\ h \qquad (9)$$

This fact again points out that a structural relationship exists between the unobservables (structural parameter estimation) rather than between the observables (conventional weighted least squares parameter estimation).

The conclusion to be reached is that (7) provides a basis of comparison with other algorithms but additional techniques need to be considered to eliminate the shortcomings of this "subtraction" method. Specifically, one needs to select M and also to minimize some weighted least squares function.

The Instrumental Variable Approach

The literature contains a method for bias removal applied to structural parameter estimation, namely, the instrumental variable[1,3,4] (IV) method. In the IV method, an additional measurement is used as an "instrument" for achieving an unbiased result. No knowledge of the noise statistics is assumed. The instrumental variable should be highly correlated with X_e but not with either noise terms (N or V). The algorithm of (3) becomes (using Z as the instrumental variable)

$$\hat{h} = (Z^T\ M\ X_s)^{-1}\ Z^T\ M\ Y_s \qquad (10)$$

The expected value of $\hat{h}$ is h, assuming Z to be correlated with X_e but not with N or V. In summary, one has adjusted the algorithm to obtain an unbiased estimate. The variance of the estimation error follows easily

$$\begin{aligned} P &= E\{(\hat{h} - h)(h - \hat{h})^T\} \\ &= (Z^T M X_e)^{-1}(Z^T M R M Z)(Z^T M X_e)^{-1} \end{aligned} \tag{11}$$

Since one can hopefully make Z highly correlated with X_e[4], then selecting $M = R^{-1}$ results in

$$P = (X_e^T M X_e)^{-1} \tag{12}$$

which is quite similar to conventional weighted least squares. One has selected M, but still one must force Z to approach X_e and one has not necessarily minimized J from (2).

New Approach

Let us consider a new approach to achieving unbiased structural parameter estimates. This approach follows directly from a weighted least squares minimization problem and the weight selection is well defined. Note that (1) can be written in two ways.

$$Y_e = X_e h \tag{13a}$$

$$Y_e = H \chi_e \tag{13b}$$

In (13a), h is an n x 1 column vector and X_e is an m x n matrix (m could equal 1, in which case X_e would be a row vector). In (13b), all the different measurements contained in X_e are reformed into a k x 1 column vector χ_e with $k \leq mn$. H is therefore an m x k matrix whose elements are formed from the n x 1 matrix h. Both (13a) and (13b) are equivalent. In view of (13a) and (13b) let us write the sensor equation (8) in two ways..

$$X_s = X_e + N \tag{14a}$$

$$\chi_s = \chi_s + n \tag{14b}$$

Consider the objective function

$$J = \hat{n}^T M_1 \hat{n} + [Y_s - (X_s - \hat{N})\hat{h}]^T M_2 [Y_s - (X_s - \hat{N})\hat{h}]$$
$$+ [Y_s - \hat{H} X_s + \hat{H} \hat{n}]^T M_2 [Y_s - \hat{H} X_s + \hat{H} \hat{n}] \tag{15}$$

At this point, if one were to obtain $\frac{\partial J}{\partial \hat{h}}$ and $\frac{\partial J}{\partial \hat{n}}$ one would obtain non-linear equations since $\hat{N}$ depends on n and $\hat{H}$ depends on h. Let us linearize the partial derivatives by considering $\hat{N}$ and $\hat{H}$ in (15) to be constants determined from the last estimates of $\hat{n}$ and $\hat{h}$. Solving the partial derivatives, the (n + 1) estimate is

$$\hat{h}_{n+1} = [(X_s^T - \hat{N}_n^T) M_2 (X_s - \hat{N}_n)]^{-1} (X_s^T - \hat{N}_n^T) M_2 Y_s \tag{16a}$$

$$\hat{n}_{n+1} = [M_1 + \hat{H}_{n+1}^T M_2 \hat{H}_{n+1}]^{-1} [\hat{H}_{n+1}^T M_2 \hat{H}_{n+1} X_s - \hat{H}_{n+1}^T M_2 Y_s] \tag{16b}$$

The algorithm of (16) is, therefore, iterative and begins with an initial estimate of N. This initial estimate could be zero, hence the process begins with $\hat{N}_o = 0$. As more data becomes available, the process can also be sequential. In fact, the matrices of (16) can easily be defined to contain submatrices thereby considering the case where k sets of data have been taken. For the repeated data case, algorithms such as (3) and (16) can be written as summations over k in terms of the corresponding submatrices contained in X_s, Y_s, etc.

It can be shown the $\hat{h}_{n+1}$ is an unbiased estimate of h if the weighting matrices are chosen as follows.

$$M_1 = S_1^{-1}$$
$$M_2 = R^{-1} \tag{17}$$

where S_1 is the covariance matrix for N and R is the covariance matrix for V. Note that complete knowledge of X_e implies that $S_1 \to 0$. In this case $M_1 \to \infty$ and $\hat{n}_{n+1} \to 0$ in (16b). Hence $\hat{N}$ approaches zero and the algorithm of (16a) becomes (3), the algorithm for conventional least squares estimation. On the other hand, complete ignorance of X_e imples that $S_1 \to \infty$. In this case $M_1 \to 0$ and (16b) can be written

$$\hat{n}_{n+1} = X_s - (\text{matrix}) \times Y_s \qquad (17)$$

Substituting (17) into (16a), we find that $\hat{h}_{n+1}$ is found by completely ignoring the sensed values X_s. The behavior at these two extremes is quite reasonable.

In summary, the algorithm of (16) follows directly from a weighted least squares minimization problem, the weight selection is defined, and the estimator provides unbiased operation in dealing with structural parameter estimation. This scheme may be referred to a linearized iterative weighted least squares technique (abbreviated LITWELS). This technique and the mnemonic are the work of the author.

General Problem to be Treated

The problem introduced in this report concerns estimation of a constant parameter. A more general problem may be solved in quite the same manner. In the general problem, the parameters may be time varying of the form

$$h = \phi h_o + \Gamma W \qquad (18)$$

$$Y_e = X_e h$$

where W is a noise of zero mean with covariance Q. When X_e is known exactly, then one obtains a discrete Kalman filter from a properly chosen weighed least squares minimization problem. When X_e is known with uncertainty, then the discrete Kalman filter yields a biased solution, but the instrumental variable approach[4] or the LITWELS approach can easily be extended to yield an unbiased solution.

Since the weighting matrices are chosen as the inverses of certain covariance matrices, methods must be chosen to estimate the covariance matrices when the noise statistics are unknown. Estimates of this type follow from proper manipulation of the objective function. Both the extension of the LITWELS technique to the time varying parameters case and the estimation of unknown noise statistics are future goals of this study.

References

1. M. G. Kendall and A. Stuart, The Advanced Theory of Statistics, Volume 2, C. Griffin and Company, LTD, 1961, pages 397-408.

2. R. L. Anderson and T. A. Bankroft, Statistical Theory in Research, McGraw Hill Book Company, 1952, Chapter 14.

3. R. E. Andeen and P. P. Shipley. Digital Adaptive Flight Control Systems for Aerospace Vehicles, AIAA Journal, Vol. 1, No. 5, May 1963.

4. P. C. Young, An Instrumental Variable Method For Real-Time Identification of a Noisy Process, Automatica, Vol. 6, pages 271-289, 1970.

ON THE SEQUENTIAL SEARCH FOR THE MAXIMUM OF AN UNKNOWN FUNCTION

by S. Yakowitz

I. INTRODUCTION

Many problems arising in engineering and operations research contexts have the following structure: The decision maker is provided with a class $\mathcal{F}$ of functions, whose common domain, $\mathcal{X}$ is specified. Some mechanism selects a function f from $\mathcal{F}$. The decision maker is not informed of this choice. He would like somehow to find a point $x^* \in \mathcal{X}$ at which f assumes its maximum value (denoted by $||f||$). Toward this end, the decision maker may sequentially and without constraint select elements $x_1, x_2, \ldots$ from $\mathcal{X}$. Upon choosing x_n, he is informed of the value $f(x_n)$. Thus the decision maker may come to learn certain features of f. Any (perhaps randomized) strategy for choosing x_n on the basis of the sequence of pairs $\{(x_j, f(x_j))\}$ will be termed a search procedure. The problem of finding a search procedure S under which, for all $f \in \mathcal{F}$, $\{f(x_n)\}$ converges to $||f||$, in some specified sense, has generated a lively body of research papers, some of which will be referenced and described in the present paper.

As an example of the sort of engineering question giving rise to a search problem, suppose that an airplane is to fly with a fixed velocity. Its fuel efficiency will then be a function of the carburation setting. If x is the relative mixture of fuel and air, and f(x) the associated fuel consumption required to maintain the aircraft's velocity, then the framework for a search problem is present. For this problem, X may be taken to be the unit interval and F, perhaps, may be considered to be the set of continuous functions on the unit interval.

Under certain restrictions on F and X, effective search procedures have been revealed. The most publicized of these is the "gradient method" which, in its simplest form, determines x_{j+1} from x_j by estimating the gradient ∇f of f at x_j (by difference approximations derived from local samples) and then setting $x_{j+1} = x_j + \lambda \nabla f(x_j)$. λ is chosen from heuristic considerations and may vary as the process evolves. If the functions of F are concave or at least unimodal and X is bounded and sufficiently regular, the gradient method will provide a Cauchy sequence $\{f(x_j)\}$ converging to $||f||$. Hadley's book Nonlinear and Dynamic Programming [1] devotes a nicely written chapter to the gradient method and its variations. The review paper by Spang [2] has an extensive bibliography on the gradient method, more recent methods of which are described in the book by Osborn and Kowilak [3].

J. Kiefer [4,5] has published interesting analyses for the case that X is a bounded interval in the real line. In particular, under the search procedure he proposes, in n trials (the number n must be specified in advance) the point x^* at which $f(x^*) = ||f||$ can be located within a distance of $1/L_n$, L_n being the nth Fibonacci number, when F is the set of unimodal functions on 0,1 . Further, the search procedure is minimax in the sense that no non-randomized strategies can improve on this operating point error uniformly in F. Bellman and Dreyfus [6] devote a chapter to this optimization approach. To this writer's knowledge, an analogous search which also possesses the minimax property has yet to be revealed for multi-dimensional X.

An intriguing search model (which is slightly closer to the path to be followed here in that probabilistic ideas are prominent and multi-modal functions are included in F) was proposed by H. Kushner [7,8] who supposed f to be a sample function from a Brownian motion process on a bounded linear interval, X. An advantage to this viewpoint is that, in addition to including multi-modal functions, ideas from Wiener prediction theory can be brought to bear on the problem of designing an optimal search procedure. Kushner points out that numerical evaluation of the optimal procedure is computationally prohibitive, but provides a search procedure under which $\lim_{n\to\infty} 1/n \sum_{i=1}^{n} f(x_i) = ||f||$, almost surely.

The research reported in this paper follows an approach sketched by S. Brooks [9]. Presumably, Brooks took X to be a finite set, and the loss associated with the function $f \in F$ and operating point $x \in X$ to be

$$L(x,f) = \text{"proportion" of points } x' \in X \text{ such that}$$
$$f(x') > f(x).$$

Then, given any positive numbers c, d, a smallest number N is readily calculated such that if $x_1, x_2 \ldots x_N$ are selected from X by a randomization which gives equal weight to each element of X, for any real-valued function f,

$$P[\max_{1<i<n} L(x_i,f) > c] < d, \qquad \text{for } n > N.$$

Brooks, as well as Kushner, consider the possibility that the measurements $\{f(x_i)\}$ may be corrupted by additive noise. These considerations will be detailed, along with a brief review of "stochastic approximation" in a later section (Section 4) of this paper.

Let us summarize the results of this paper. $\mathcal{F}$ will, in all our studies, at least include the set of continuous functions on $\mathcal{X}$, which, for expository reasons, will be the unit interval. Generally, capital letters denote random variables and lower case letters an observation of the variable designated by the capitalization. Section 2 reveals two random search procedures; the first of these achieves almost sure convergence of $1/n \sum_{i=1}^{n} f(X_i)$ to $||f||$ for each $f\epsilon\mathcal{F}$, and the second yields a random sequence $\{f(X_i)\}$ which converges in probability to $||f||$. Section 2 concludes with a theorem on the non-existence of a search procedure under which $f(X_n) \rightarrow ||f||$ almost surely for all continuous f, and a theorem on the impossibility of bounding the rate of convergence in probability.

Section 3 reopens and extends the research path suggested by Brooks [9]. Where Brooks defines the loss associated with $f\epsilon\mathcal{F}$ and operating point $x\epsilon\mathcal{X}$ by "proportion" of $x'\epsilon\mathcal{X}$ such that $f(x')>f(x)$, we define the loss to be

$$L(x,f) = \text{Lebesgue measure} \quad x': \{f(x') > f(x)\}.$$

It will be verified that this retains the important feature in Brooks' study that, for any positive numbers c and d, one may compute in advance of making measurements, how many measurements N are required so that, for any $f\epsilon\mathcal{F}$, $n \geq N$,

$$P[L(X_{n^*}, f) > c] < d, \tag{1.1}$$

n^* being the random element i, $1 < i < n$, which maximizes the measurement $f(X_i)$. Further, random searches S_1 and S_2 and numbers N_1 and N_2 are described such that, for any $f \epsilon F$, under S_1,

$$P[\sup_{n>N_1} 1/n \sum_{i=1}^{n} L(X_i, f) > c] < d \tag{1.2}$$

and under S_2

$$P[L(X_n, f) > c] < d \quad \text{for all } n > N_2. \tag{1.3}$$

At the close of Section 2, we show that under certain mild restrictions, with increasing n the random variable n $L(X_{n^*}, f)$ converges weakly to the exponential variable with parameter 1. In this statement, n^* has the same meaning as given in connection with equation (1.1).

Section 3 studies the case that the measurements $\{f(x_i)\}$ are corrupted by independent, identically distributed additive noise, which is assumed not to depend on f. With no further assumptions on the noise process, we reveal a search procedure under which the average operating loss, $1/n \sum_{i=1}^{n} L(X_i, f)$, converges in probability to 0 for every Lebesgue-measureable function f; in the noisy case, however, no lower bounds for the rate of this convergence have been discovered. If the noise distribution is known, the previously mentioned convergence is obtainable even if the noise distribution depends on the operationg point x. We compare this noisy-measurement problem and the results obtained to the class of problems which are known to yield to the method of stochastic approximation; also related results due to Kushner are menttioned.

The concluding section suggests how the preceding theory can be extended to unbounded and multi-dimension X and mentions a few implications of these studies.

II. ON THE EXISTENCE OF CONVERGENT SEARCHES

To recapitulate certain remarks made in the previous section, the situation with respect to convergence (or equivalently, almost certain convergence) of $\{f(x_i)\}$ to $||f||$ is that this problem has been solved only in certain weak senses. Gradient techniques as well as Fibonacci searches require at least that f be unimodal. The only other principal result that this author has uncovered in the literature is that if f is a sample function of a given Wiener process on X, then there is a search procedure achieving convergence almost surely of $\sum_{i=1}^{n} f(x_i)/n$ to $||f||$. (For brevity, let us refer to this analysis [6] as "Kushner theory"). The weakness of the situation cited above are grevious. First, the available models are too restrictive; Many classes of important criterion functions are excluded, or (as in the Kushner theory case) a structure is imposed that will not serve as a natural model for many "real-world" phenomena. Second, engineers and other practitioners of control theory are, or should be, desirous of having some means of computing how many observations are required to achieve a certain performance level. Such a statement would go something like: "Given any δ, $\epsilon > 0$, under search procedure S, a number $N(\delta, \epsilon)$ may be computed such that $P[G(||f|| - f(X_n)) > \epsilon] < \delta$ for all $n > N(\delta, \mathcal{F}\epsilon)$, $f \epsilon \mathcal{F}$." $G(\cdot)$ would be some monotonic function of $||f|| - f(X_n)$ such as $[(f(x_n) - ||f||)/||f||)]^2$. No such results, except in extremely exclusive settings, have been revealed and generally speaking, there is no way for the engineer to estimate in advance the quality of performance obtainable in a finite number of observations.

This author has the opinion that for applications (i.e. when only finitely many observations can be made), convergence in probability is fully as valuable as convergence almost surely, and both convergences are essentially worthless unless associated bounds can be derived. However, if the reader is willing to attach value to convergence unaccompanied by bounds, then the following theorems may be of interest in that they describe search procedures simultaneously as effective (in terms of convergence achievable if the distinction between convergence in probability and almost surely is ignored) as gradient methods, Fibonacci search, or searches in Kushner theory, but which are valid under a much more general setting.

Theorem 1: Let F be the set of bounded, piecewise-continuous functions on [0, 1]. One may compute a search procedure S_1 under which, for every $\epsilon > 0$

$$\lim_{n\to\infty} P[f(X_n) < ||f|| - \epsilon] = 0$$

for every $f\epsilon F$.

Theorem 2: Under the conditions of Theorem 1, one may compute a search procedure S_2 under which

$$\lim_{n\to\infty} (1/n \sum_{i=1}^{n} f(X_i) = ||f||$$

almost surely for every $f\epsilon F$.

Theorems 1 and 2 are trivial consequences of Theorems 5 and 6, in the proofs of which are described search procedures S_1 and S_2.

Theoreticians might appreciate our observing that under a class as large as F above, in Theorem 1 weak convergence cannot be strengthened to strong convergence, as we now demonstrate.

Theorem 3: If F is the space of continuous functions on $[0, 1]$, there is no search procedure under which $f(X_n) \to ||f||$ almost surely for all $f \epsilon F$.

PROOF: Let f be any continuous function taking its maximizing value $||f||$ at some unique point x^* interior to $[0,1]$, and suppose S is some search procedure under which $f(X_n) \to ||f||$ almost surely. Let a be any positive number less than 1/2. Define A_n to be the event that $x^* - a < x_i < x^* + a$ for all $i > n$. Notice that A_n is an increasing sequence, and as x^* is a unique maximizing point of f, $\lim_{n\to\infty} P[A_n] = 1$. Thus for some integer N, $P[A_N] > 0$ and as there are infinitely many disjoint non-degenerate intervals in X but outside the interval $[x^* - a, x^* + a]$, there must be some interval I such that the event "I is not sampled at all" has positive probability under the process determined by S on f. B will denote this event. Let f' be a function identical to f on I^c and assuming a maximum $||f'|| > ||f||$. (The maximizing point or points must be interior to I.) Recall that search procedures are constrained to depend on the function being searched only through values actually sampled. Consequently, S on f, given B and S on f' given B are the same process. As $f(x_i) \to ||f||$ almost surely, and B has positive probability then, given B, $f'(x_i) \to ||f|| < ||f'||$. One concludes that under S, $P[f'(x_i) \to ||f'||] \leq 1 - P[B] < 1$.

As we have mentioned, for application of an optimization procedure, it seems to us highly desireable that, within the mathematical framework of the procedure, there be some way of assessing what can be done in a finite number of iterations. The theorem to follow suggests that such estimates will not be available under the loss criterion "$||f|| - f(x_n)$".

Theorem 4: If F is the space of continuous functions on [0,1], no search procedure exists such that for every $c,d > 0$, there is an integer $N(c,d)$ for which, if $i > N(c,d)$,

$$P[(||f|| - f(X_i))/||f|| > c] < d$$

for every $f \epsilon F$.

PROOF: Let S be any search procedure, c and d any positive numbers less than 1, f any function in F, and N any positive integer. Define I to be some interval in [0,1] such that, under the random sequence induced by S on f, the event (call it B) "I is sampled by time N" has probability less than d. f' is any function in F which agrees with f on I^c an has a maximum which satisfies the inequality

$$(||f'|| - ||f||)/||f'|| > c.$$

Then, given B, $f'(x_i) \leq ||f||$ for $1 \leq i \leq N$, and consequently, for the process induced by S on f'

$$P[||f'|| - f'(X_n)/||f'|| > c] \geq P[B] > d \quad \text{for } n \leq N. \qquad (2.1)$$

As N and S are arbitrarily chosen, (2.1) implies the theorem.

The preceding development gives ample evidence for the assertion that if one wishes a search procedure having convergence bounds uniform on the set of continuous functions, it is necessary to consider a loss criterion different from monotonic functions of $||f|| - f(X_n)$. The next section suggests such an alternative for which uniform bounds are revealed.

III. PROPERTIES OF THE MEASURE OF THE DOMAIN OF IMPROVEMENT

We seek to overcome what we regard as the greatest weakness in the existing theory of search procedures (which was described in the previous section), namely, in the class F of continuous functions on [0,1], under no search procedure can bounds on the rate of convergence of $f(x_i)$ to $||f||$, or on the rate of convergence of $1/n \sum_{i=1}^{n} f(x_i)$ to $||f||$ be established which are uniform on F. The practical consequence of this weakness is that the experimenter cannot estimate the level of performance obtainable in a finite number of search iterations. Our approach to overcoming these difficulties is to redefine the search problem by proposing a different (but, hopefully not unreasonable) criterion of goodness.

Associated with each operating point $x \epsilon F$ and criterion function $f \epsilon F$ is the set $A(x,f) = \{y:f(y) > f(x)\}$, which is here called the <u>domain of improvement</u> (of f over f(x)). We propose, as a loss function for search problems, the Lebesgue measure, $m(A(x,f))$, of $A(x,f)$. Thus $L(x,f) = m(A(x,f)) = m[f > f(x)]$. Note that for every continuous function f, the loss function L(x,f) imposes the same partial ordering on $\mathcal{X}$ as does $||f|| - f(x)$ (i.e. $L(x,f) < L(y,f)$ if and only if, $||f|| - f(x) < ||f|| - f(y)$). Obviously, then, $L(x_n,f) \rightarrow 0$ and $||f|| - f(x_n) \rightarrow 0$ are equivalent statements. Thus, in an important sense, the classical loss function and the measure of the domain of convergence are equivalent.

We remind the reader that in Section I, with respect to a fixed function f, for a sequence $\{x_i\}_{i=1}^{n}$ we defined n* to be any subscript $m(1 \leq m \leq n)$ such that

$$f(x_m) = \max \ f(x_i): \ 1 < i < n .$$

The strength of the results on search procedures, under the loss $L(x,f)$ stem from the fact that if $\mathcal{X} = [0,1]$ and $\{X_i\}_{i=1}^{n}$ is a sequence of independent random variables uniformly distributed on $\mathcal{X}$ then $L(X_{n*},f)$ has a distribution which is essentially independent of $f\epsilon\mathcal{F}$(or, more accurately, has a "worst case" in $\mathcal{F}$).

Theorem 5: Let $\mathcal{F}$ be the set of Lebesgue-measureable functions on [0,1]. For any integer n and number a in the open unit interval,

$$P[L(X_{n*},f) \geq a] \leq (1-a)^n$$

for every $f\epsilon\mathcal{F}$.

PROOF: Let $t' = \inf\{t:m[f > t] \geq a\}$. As $m[f > t]$ is continuous from above, $m[f > t'] \geq a$. Also, since Lebesgue measure and the uniform probability coincide on Borel subsets of $\mathcal{X}$,

$$1 - a \geq m[f \leq t'] = P[f(X_i) \leq t'] = P[L(X_i,f) \geq a \ , \ 1 \leq i \leq n.$$

In order that $f(X_{n*}) \leq t'$, we must have that $f(X_i) \leq t'$, $1 \leq i \leq n$.

Therefore,

$$P[L(X_{n*},f) > a] = P[f(X_i) \le t'; 1 \le i \le n] = \prod_{i=1}^{n} P[f(X_i) \le t'] \le (1-a)^n.$$

If m[f > t] is continuous, then for each a in the unit interval there is a t' such that m[f > t'] = a, and thus the bound described in Theorem 5 cannot be improved upon. For example, if $f(x) = x$, $(x \in \mathcal{X})$, then

$$P[L(X_{n*},f) > a] = (1-a)^n. \qquad (3.1)$$

In what follows, M_n will denote the random variable $L(X_{n*},f)$ determined by equation (3.1). That is, M_n is the random variable having the cumulative distribution function $F_n(x) = 1 - (1-x)^n$, $(0 \le x \le 1)$. For several numbers A and D, Table I gives the maximum number of observations N requires so that $P[M_N > A] < D$. Also in this section, $\mathcal{F}$ will denote the set of Lebesgue-measureable functions on $[0,1](= \mathcal{X})$.

A D	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
.05	59	29	19	14	11	9	7	6	6	5
.10	45	22τ	15	11	9	7	6	5	4	4
.15	37	19	12	9	7	6	5	4	4	3
.20	32	16	10	8	6	5	4	4	3	3
.25	28	14	9	7	5	4	4	3	3	2
.30	24	12	8	6	5	4	3	3	3	2
.35	21	10	7	5	4	3	3	3	2	2
.40	18	9	6	5	4	3	3	2	2	2
.45	16	8	5	4	3	3	2	2	2	2
.50	14	7	5	4	3	2	2	2	2	2

A D	.005	.010	.015	.020	.025	.030	.035	.040	.045	.050
.005	1058	528	351	263	210	174	149	130	116	104
.010	919	459	305	228	182	152	130	113	101	90
.015	838	418	278	208	166	138	118	103	92	82
.020	781	390	259	194	156	129	110	96	85	77
.025	736	368	245	183	146	122	104	91	81	72
.030	700	349	233	174	139	116	99	86	77	69
.035	669	334	222	166	133	111	95	83	73	66
.040	643	321	213	160	128	106	91	79	70	63
.045	619	309	206	154	123	102	88	76	68	61
.050	598	299	199	149	119	99	85	74	66	59

TABLE I

MINIMUM N SUCH THAT P[M_N > A]< D

Let us proceed to our goal of revealing search procedures achieving bounded convergence to optimal performance.

Theorem 6: One may compute a search procedure S_1 under which, for any positive numbers c and d, a number N(c,d) may be found for which

$$P[\sup_{n>N(c,d)} 1/n \sum_{i=1}^{n} L(X_i,f) > c] < d$$

for every $f \epsilon F$.

PROOF: Let $\{n(i)\}_{i=1}^{\infty}$ be a sequence of numbers such that n(1) = 1 and i/n(i) converges to 0 monotonically (e.g. $\{2^{i-1}\}$). By theorem 5, we may compute a number N' such that

$$P[M_N' > c/2] < d.$$

Also, we may find a number N" greater than N' such that

$$(c/2)[(n_{N''} - N'')/n_{N''}] + 1[(n_{N'} + N'')/n_{N''}] < c.$$

Search procedure S_1 requires that X be sampled independently and uniformly at times $t = n_j$ $(j = 1, 2, ...)$, and for $t \neq n_j$, x_t is chosen to be the best value in the sequence $\{X_{n(j)}\}$ sampled thus far: $f(x_t) = \max \{f(X_v): v \leq t\}$. Thus evidently $f(x_t)$, $t \notin \{n_{(j)}\}$ is monotonically increasing in t. Observe that from the choice of N' and the definition of S_1,

$$P[L(X_{n(N')},f) > c/2] < d.$$

Let Q be the event (with reference to the process determined by S_1 on f) that $L(x_{n(N')}f) \leq c/2$. If Q occurs, then by the choice of N" (and observation that $L(x,f) \leq 1$, always)

$$\sup_{n>N''} \sum_{i=1}^{n} 1/n\ L(Xi,f) \le c.$$

In summary,

$$P[\sup_{n>N''} 1/n \sum_{i=1}^{n} [L(X_i,f) > c] \le P[Q^c] < d,$$

and consequently the theorem is proved, with the understanding that N" suffices for N(c,d).

Theorem 7: One may compute a search procedure S_2, under which, for any positive numbers c and d, a number N(c,d) may be found for which

$$P[L(X_n,f) > c] < d$$

for all n > N(c,d) and all $f\in\mathcal{F}$.

PROOF: Let $\{n(j)\}$ be a sparse sequence as in the proof of Theorem 6. From this we construct a random sequence $\{N(j)\}$ where N(j) has the sample space $\{n(j), n(j) + 1, n(j) + 2, \ldots, n(j+1) - 1\}$ and is chosen by the randomization which assigns equal probability to each element of this sample space. S_2 is the search procedure which samples $\mathcal{X}$ independently and uniformly at times in $\{N(j)\}$. At other times, x_t is chosen to be the best operating point thus far sampled. The condition imposed on $\{n(j)\}$ that j/n(j) converge monotonically to 0 as j tends to infinity ensures us that a number N' can be found such that $P[N(j) = n] < d/2$ for all j > N', all integers n. From Theorem 5, a number N" may be found such that $P[M_{N''} > c] < d/2$. If $k = \{\max N',N''\}$ then for n > n(k)

$$P[L(X_n,f) > c] \le P[M_{N''} > c] + P[n\in\{N(j)\}] < d.$$

Now let us turn our attention to the question of how fast M_n converges, as n increases. In the spirit of the central limit problem, we seek a sequence $\{g(n)\}$ such that $g(n)M_n$ converges to a limiting random variable other than the unitary variable, and we wish to find also what this limiting variable is. If we are able to resolve this problem, then, heuristically speaking, $1/g(n)$ will be the convergence rate of M_n. In answer to these questions, we will find that M_n converges to the exponential variable at the rate of $1/n$. Also, we will be able to bound the error induced by replacing the cumulative distribution function $F_n(x/n)$ of nM_n by its limit distribution, $1 - e^{-x} \equiv F(x)$.

Theorem 8: $\{nM_n\}$ converges weakly to the exponential variable with parameter 1.

PROOF:

$$F_n(x/n) = 1 - (1-x/n)^n = P[nM_n \leq x].$$

Thus obviously,

$$\lim_{n\to\infty} P[nM_n \leq x] = \lim F_n(x/n) = \lim 1 - (1-x/n)^n = 1 - e^{-x}$$

$$\equiv F(x), \qquad x > 0.$$

By using Taylor's formula with remainder on the logarithm of $e^{-x}/(1-x/n)^n$, one may verify that for $x > 0$, $n = 1, 2, \ldots,$

$$\exp(-x^2/2n) < [1 - F(x)]/[1 - F_n(x/n)] < \exp(-x^2/2n + x^3/6n^2).$$

IV. SEQUENTIAL SEARCH USING NOISY MEASUREMENTS

To the structure of the sequential search problem considered in earlier sections, this section appends the possibility that upon selection of operating point x_n at the nth search iteration, the decision-maker observes

$$f(x_n) + Z_n \tag{4.1}$$

where $\{Z_n\}$ is a sequence of independent random variables (rv's). To begin with, we will assume the Z_n's to be identically distributed, but ways in which this restriction may be relaxed will be mentioned. Physically, $f(x_n) + Z_n$ may be regarded as arising from a noisy meter which measures $f(x_n)$. The above restrictions imply that the noise characteristics of the meter are independent of previous measurements as well as the magnitude of the quantity being measured. A search procedure S_3 will be revealed under which $1/n \sum_{j=1}^{n} L(x_j,f)$ converges to 0 in probability for all Lebesgue-measureable functions f, regardless of the common distribution of the Z_i's. In contrast to the noiseless case, a lower bound to the rate of convergence is not available. Connection of our study here to related results in the theory of stochastic approximation and Kushner theory will be mentioned.

In the theorem to follow, the restriction that the Z_n's be independent random variables identically distributed as Z is assumed to be in force. "Noisy measurements" refer to observations of the form (4.1) (in contrast to $f(x_n)$, which is considered a "noiseless measurement"). As in Section 3, $\mathcal{F}$ is the set of Lebesgue-measureable functions on $\mathcal{X}$, the unit interval.

Theorem 9: One may compute a search procedure S_3 on noisy measurements under which

$$1/n \sum_{j=1}^{n} L(x_j,f) \to 0 \quad \text{almost surely}$$

for all noise distributions Z and all $f \in F$ for which there is a sequence (w_n) such that $L(w_n,f) > 0$, $n = 1, 2, \ldots,$ and $\lim_{n\to\infty} L(w_n,f) = 0$.

Remark: For piecewise continuous functions f, this last restriction is satisfied if f does not assume its maximum on a plateau.

PROOF: The description of the search procedure S_3 uses the following notation: $\{u(n)\}$ is an observation of a sequence of independent rv's $\{U(n)\}$ uniformly distributed on $\mathcal{X}$. $R_{N(j)}$ denotes the empiric distribution function constructed from the observations which, during the course of the search, have been made at u(j), j = 1, 2, ... (An empiric distribution function F_n constructed from any sequence $\{x_i\}\sum_{i=1}^{n}$ of n real numbers is the cumulative distribution function determined by the expression

$$nF_n(x) = \text{number of elements } x_j \text{ of } \{x_i\}\sum_{i=1}^{n} \text{ such that } x_j \le x.$$

$F_{u(j)}$ is the cumulative distribution function (cdf) for the rv $f(u(j)) + Z$; i.e.,

$$F_{u(j)}(z) = F_Z(z + f(u(j))), \qquad \text{for every real } z.$$

More generally, F_x is the cdf of $f(x) + Z$. If $H(x)$ is any real function, the norm $||H|| = \sup_{x\in\mathcal{X}} H(x)$. $\{K(v)\}$ is a sequence of integers such that if $n > K(v)$, then for any cdf F, and empiric distribution function F_n constructed from n independent observations distributed as F,

$$P[||F - F_n|| \geq 1/v] < 2^{-v}/v. \qquad (4.2)$$

Massey [10] gives an algorithm capable of computing a minimum such number K(v). {M(v)} is a sequence computed inductively by the following rule:

$$M(2) = 1.$$

$$M(v) = M(v - 1) + A(v) + v\,K(v), \qquad v > 2$$

where A(v) is some positive integer such that

$$M(v-1) + v\,K(v) + (v + 1)\,K(v + 1)\;/A(v) < 1/v.$$

Having described {K(v)} and {M(v)}, we are in a position to reveal the search procedure S_3.

Step 1:

For each iteration v, v = 2, 3, ..., of these steps 1-3, the points $\{x_n\}_{n=M(v)}^{M(v)+vK(v)}$ are chosen, at each n, from the set of points {u(j): j = 1, 2, ..., v}, so that each u(j) is sampled K(v) times. Therefore, by time M(v) + vK(v),

$$P[||F_{N(j)} - F_{u(j)}|| \leq 1/v,\ j = 1, 2, \ldots, v] > 1 - 2^{-v}. \qquad (4.3)$$

Step 2:

At time M(v) + vK(v), a positive integer $v^* \leq v$ is selected such that for every real number z,

$$F_{N(v^*)}(z) > F_{N(k)}(z) - 2/v \quad \text{for } 1 \leq k \leq v. \qquad (4.4)$$

If no such v* can be selected, v* is chosen arbitrarily.

Step 3:

At times n, $M(v) + vK(v) < n < M(v + 1)$, $x_n = u(v^*)$. At time $M(v + 1)$, step 1 is repeated, with v increased by 1. Toward outlining a proof that S_3, as just described, possess the property asserted in the theorem, it is necessary to recognize that with probability 1, (4.4) will hold for all but finitely many v. For demonstration of this, let u(v') be any number such that

$$f(u(V')) = \max_{1 \le j \le v} f(u(j)).$$

Then for all z and all $i \le v$,

$$F_{u(v')}(z) = F_Z(z + f(u(v'))) \ge F_{u(i)}(z) = F_Z(z + f(u(i))).$$

The event (which will be denoted by B(v)) that

$$||F_{N(j)} - F_{u(j)}|| \le 1/v, \qquad 1 \le j \le v \tag{4.5}$$

implies, by the triangle inequality, that for $j \le v$,

$$F_{N(v')}(z) > F_{N(j)}(z) - 2/v, \qquad \text{all real } z$$

and thus (4.4) holds with $v^* = v'$. Note that by construction of $\{k(v)\}$,

$$\sum_{v=2}^{\infty} P(B(v)^c) < \sum_{v=2}^{\infty} 2^{-v} < \infty$$

and consequently, by the Borel-Cantelli lemma, B(v) occurs for but finitely many v, concluding our assertion that for all but finitely many v, concluding our assertion that for all but finitely many v,

v^* can be picked to satisfy (4.4). We will hereafter assume without comment that $v*$ always has the property (4.4). As our only concern is with limit theorems, this assumption will not lead us astray.

The completion of the proof that S_3 leads to the convergence of $1/n \sum_{i=1}^{n} L(x_i,f)$ to 0 is at hand. By the choice of $M(v)$ and $A(v)$, we have that at all times Q during the vth iteration of steps 1-3 ($v > 2$) that

$$\text{Number of Observations } x_i,\ 1 < i < Q, \text{ taken at } (v-1)^* \text{ or } v^*]/Q > (v-1)/v,$$

and thus for all $n > M(3)$,

$$1/n \sum_{i=1}^{n} L(x_i,f)$$

$$< 1/v + ((v-1)/v) \max \{L(u(v*),f),\ L(u((v-1)^*),f)\}. \qquad (4.6)$$

The proof is completed by showing that almost surely,

$$L(u(v^*),f) \to 0.$$

Let x' be any point in X such that $L(x',f) > 0$. Then certainly, some element $u(h)$ in an observation of $\{U(v)\}$ gives $f(u(h)) > f(x')$. If H is a number such that

$$6/H < \| F_{x'} - F_{u(h)} \|,$$

then for all $v > \max \{H,h\}$, if $f(u(j)) \leq f(x')$,

$$F_{N(v^*)}(z) \geq F_{u(h)}(z) - 2/v > F_{u(j)}(z) + 6/H - 2/v$$

$$> F_{N(j)}(z) + 6/H - 4/v > F_{N(j)}(z) + 2/v, \text{ (all real z)},$$

which implies that j cannot be chosen to satisfy (4.4) for v^*. From this we deduce that

$$\lim\sup L(u(v*),f) \leq L(x',f). \qquad (4.7)$$

Let $\{w_n\}$ be a sequence, as in the hypothesis of the theorem, such that $L(w_n,f) > 0$ and $\{L(w_n,f)\} \rightarrow 0$. Then (4.7) holds almost surely simultaneously for all the w_n (in place of x') and we conclude that with probability 1,

$$\lim L(u(v^*),f) \leq \inf_n L(w_n,f) = 0.$$

Theorem 9: <u>One may compute a search procedure</u> $S_3{}'$ <u>on noisy measurements under which</u>

$$(1/n) \sum_{i=1}^{n} L(X_i,f) \rightarrow 0 \quad \text{in probability}$$

<u>for all noise distributions and all</u> $f\epsilon F$.

PROOF: $S_3{}'$ differs from S_3 only in step 2, where for $S_3{}'$ the restriction is made that v be the <u>greatest</u> positive integer v such that for every real number z,

$$F_{N(v^*)}(z) > F_{N(k)}(z) - 2/v, \quad 1 \leq k \leq v. \qquad (4.8)$$

Observe that $S_3{}'$ is a version of S_3, and consequently it achieves convergence under the hypothesis of the preceding theorem.

In the absence of a sequence $\{w_n\}$ as hypothesized in the previous theorem, there is a number t' such that

$$m\ f > t' = 0 \quad \text{and} \quad m\ f = t' > 0. \qquad (4.9)$$

(As in Section 3, we use the abbreviation $m[f > b]$ to denote the Lebesgue measure of the domain of improvement $\{x:f(x) > b\}$). We use the notation of the proof to the preceding theorem. Let h be an integer (surely there is one) such that $f(u(h)) = t'$. Then for $v \geq h$, under S_3', v becomes v^* by virtue of one of the events $A(v)$ or $B(v)$ (in the sigma-field of the process determined by S_3 and f) occurring:

$$\begin{array}{lll} & A(v): & f(U(v)) = t'. \\ & B(v): & B(n) = B_1(v)\ \ B_2(v). \\ \text{where} & & \\ & B_1(v): & t' > f(u(v)) \geq t' - a(v) \\ \text{and} & & \\ & B_2(v): & F_{N(v)} \text{ satisfies} \end{array} \tag{4.8}$$

Here $\qquad a(v) = \inf \{a: ||F_{t'} - F_a|| \leq 2/v \}$.

Note that $P[A(v) \cup B(v)] \geq P[A(v)] = m[f = t']$ which is positive and independent of v. Thus under S_3', infinitely many different v are chosen as v^*. Our proof consists of showing (below) that

$$\lim_v P[B(v) | A(v) \cup B(v)] = 0 \tag{4.10}$$

Note that $A(v)$ and $B_1(v)$ are independent of $\{U(k): k \neq v\}$. Thus (4.10) implies that $\lim_v P\ f(U(v^*)) = t' = 1$, which in turn implies that $\{L(U(v^*), f)\}$ converges in probability to 0. This (in view of equation (4.6)) concludes the proof.

We proceed now to the demonstration of (4.10).

$$P[B(v) | A(v) \cup B(v)] \leq P\ B_1(v) | A(v) \cup B(v)]$$

$$= P[t' > f(U(v)) \geq t' - a(v)]/P[t' \geq f(u(v)) \geq t' - a\ (v)].$$

As $\{a(v)\}$ converges to 0 monotonically, by the continuity property of measures,

$$\lim_v P[t' > f(U(v)) \geq t' - a(v)] = 0.$$

Similarly,

$$\lim_v P[t' \geq f(U(v)) \geq t' - a(v)] = P[f(U(v)) = t'] = m[f = t'] > 0.$$

Thus $P[B_1(v) | A(v) \cup B(v)] \to 0$, which in turn implies that $P[B(v) | A(v) \cup B(v)] \to 0$.

We discuss briefly possible extensions of the preceding theory for search procedures using noisy measurements. First we mention that under the hypothesis of the preceding theorem (Theorem 9), search procedures may be devised to achieve convergence in probability of $L(X_n, f)$ to 0, for all $f \in \mathcal{F}$. One way is to choose $M(v)$ (described in the proof to Theorem 9) randomly and sufficiently sparsely.

Next we consider different assumptions about the noise process. The search procedure S_3' described in this section is effective regardless of the noise distribution F_Z. Our results cannot essentially be improved, for generally even if F_Z is specified in advance, uniform (on $\mathcal{F}$) lower bounds for convergence cannot be obtained. On the other hand, if the noise distribution is available in advance, we may let the noise depend on the operating point $x \in \mathcal{X}$. That is, if the measurement made at operating point x is the random variable $f(x) + Z(x)$, $Z(x)$ being a random variable with known cdf F^x, then there are search procedures which achieve convergence in probability of

$$\text{i)} \qquad 1/n \sum_{i=1}^{n} L(X_i, f) \to 0$$

and

$$\text{ii)} \qquad L(X_i, f) \to 0.$$

for all measureable functions f

In the notation of Theorem 8, let us sketch how convergence i) may be achieved, using an iterative search procedure. Step one begins at time M(v), $v \geq 2$. $M(2) = 1$

Step 1: Find a number δ_j such that

$$||F^{u(j)}(z) - F^{u(j)}(z + (1/v))|| > 2\delta_j, \qquad 1 \leq j \leq v.$$

Sample at u(j), $(1 \leq j \leq v)$, sufficiently many times so that

$$P[||F_{N(j)} - F|| < \delta_j, \text{ for } 1 \leq j \leq v] > 1 - 2^{-v}. \qquad (4.11)$$

In (4.11), $F_{u(j)}$ is the distribution of $f(u(j)) + Z(u(j))$. K(v) is defined to be the number of observations required to achieve (4.11).

Step 2: Define $f'_j(v)$, if possible, so that

$$||F_{N(j)} - F^{u(j)}(z + f'_j(v))|| < \delta_j. \qquad (4.12)$$

v^* is defined to be the greatest integer $k(k \leq v)$ such that

$$f'_k(v) = \max_{1\leq j\leq v} f'_j(v). \qquad (4.13)$$

Step 3: If step 1 began with the M(v)th observation, sample at $u(v^*)$ until time $M(v) + K(v) + A(v)$, where A(v) is a positive integer large enough that

$$M(v) + K(v) + K(v + 1)/A(v) < 1/v.$$

Increase v by 1 and return to step 1, setting the new M(v) to $M(v) + K(v) + A(v) + 1$.

Notice that if (4.12) and the event in (4.11) are satisfied,

$$||F^{u(j)}(z - f'_j(v)) - F^{u(j)}(z - f(u(j)))|| \le || F^{u(j)}(z - f'_j(v)) - F_{N(j)}||$$

$$+ ||F_{N(j)} - F^{u(j)}(z - f(u(j)))|| < 2\delta_j,$$

which, by choice of δ_j, implies that

$$|f(u(j)) - f'_j(v)| < 1/v. \qquad (4.14)$$

From this we see that under our search procedure,

$$P[|f(u(j)) - f'_j(v) | < 1/v, \ 1 \le j \le v] > 1 - 2^{-v}. \qquad (4.15)$$

The conclusion that $\{L(U(v\ ),f)\}$ converges in probability to 0 (and consequently so does $1/n \sum_{i=1}^{n} L(X_i,f)$) is readily derived from (4.6).

We close this section by mentioning related studies. Brooks [9] mentioned the idea of overcoming noise by repeatedly sampling at each operating point. We have stated that in Kushner theory [7] it is supposed that f is a sample function of a known Brownian motion process. It is further allowed that the measurement may be corrupted by Gaussian noise having zero mean and a known variance, which is allowed to depend on the operating point x. The framework for computing an optimal search procedure minimizing $E[(||f|| - f(X_n))^2]$ is sketched, but it is not proven that these methods yield convergence of the above expectation to 0.

Our studies are also somewhat related to the subject of "stochastic approximation," initiated by Monro and Robbins [11] and placed in an optimization setting by Kiefer and Wolfowitz [12]. A definitive survey of stochastic approximation has been written by Schmetterer [13]. Briefly, the stochastic approximation problem in determining the maximum of a regression function may be viewed as the problem of finding a search procedure yielding a sequence $\{X_i\}$ converging (either in probability or almost surely) to x^*, where x^* is the unique operating point maximizing f. The stochastic approximation setting is more general than ours in that the noise process, while (as in our studies) being independent of earlier observations, may be unknown and yet depend on x. But it is at the same time more restrictive than our theory because f must be a function which is unimodal, i.e. monotonically increasing for $x < x^*$ and monotonically decreasing for $x > x^*$. There are various other assumptions imposed on both F and the noise process; the reader is invited to consult the stochastic approximation references for details of this very deeply researched theory.

V. SUMMARY AND EXTENSIONS

It is evident that the methods of this paper can be used for bounded intervals other than the unit interval. In fact, any Lebesgue set having positive finite measure may play the role of $\mathcal{X}$. Also, no doubt the reader has noticed that while we were assuming $\mathcal{X}$ to be the unit interval, in all sections but the preceding no restriction, or even changes, are required if instead, $\mathcal{X}$ is taken to be the unit n-cube. Of course, in higher dimension, m[A] is the multi-dimensional Lebesgue measure of the set A, and with this measure, $L(x,f) \quad m[f > f(x)]$. The uniform searches in higher dimension problems are again uniform searches (in the higher dimension $\mathcal{X}$). A point where it may not be clear that the theory can be extended is in the noisy measurement problem studied in the preceding section. It is not perhaps well-known that there is a way to find K(v) if F and F_n are n-dimensional cdf's. Nevertheless, it is true that for higher dimensions, K(v) can be computed as Kiefer and Wolfowtiz have proven [14, esp. pp. 181-182]. The K(v) computed by means of the preceding reference is not close to the minimum possible K(v), and it depends on the dimension of $\mathcal{X}$. With this last exception, our theory is independent of dimension.

Extension of our theory to sets $\mathcal{X}$ which are unbounded intervals or other sets with infinite Lebesgue measure requires more adaptation. One possibility of bringing such sets into the framework of the preceding analysis is to accept in place of m, some linear measure m' (such as the Gaussian probability measure) which assigns a finite number to the real line. One then assumes that the loss L(x,f) associated with operating at point x is $m'[f>f(x)]$. Our analysis remains valid if, instead of sampling $\mathcal{X}$ uniformly, it is sampled according to a probability function P such that for some c,

$$P[A] = c\ m'[A], \quad \text{all Borel subsets } A \subset \mathcal{X}.$$

A Bayesian might want to follow this approach regardless of the Lebesgue measure of $\mathcal{X}$ in order to take advantage of a priori ideas about the location of maximizing values of f.

While on the subject of the Bayesian viewpoint, we mention that if a cost c is attached to making each observation, and a loss L' (L(x,f)), a monotonic function of L(xf), is associated with stopping the search when $x_{n*} = x$ and f is the unknown criterion function, then the optimal stopping rule is to stop after the

Tth sample, where T is the greatest integer such that

$$c < E\,[L'[(L(X_{T*},f))] \;-\; E\,[L'\,(L(X_{(T+1)*},f))]$$

We have seen that the distribution of $L(X_{n*},f)$ is independent of f if $m[\,f>t\,]$ is continuous. Thus the optimal stopping time T may be determined in advance of making measurement.

If $\mathcal{F}$ is a set of uniformly bounded measureable functions, the uniform bound M being known, and if

$$L''(x,f) \equiv \int_E f(y)dy, \qquad E = \{y:f(y) > f(x)\},$$

then $L''(x,f) < M\, m(f>f(x))$ and one sees that the preceding theory is applicable for finding search procedures under which the loss, as measured by L'', converges, in the respective senses, to zero. We are indebted to our colleague Dr. A. Wayne Wymore, for suggesting this observation.

The goal in this paper has been to delimit what can be done by sequential search procedures when the set of objective functions is rich enough to include all continuous functions. Where possible, we have sought bounds to the number of observations needed to accomplish those results that can be accomplished. This goal is more in the tradition of automata theory than numerical analysis. Toward this goal we have revealed several search procedures giving convergence (in various senses) to optimal performance. Many of these results, especially in the noisy measurement case, are believed to be new.

For particular numerical problems wherein some prior knowledge of the criterion function f is available, we expect that often heuristic considerations will yield more rapid convergence than our algorithms. The literature suggests that heuristic "creeping search" programs (e.g. Schumer and Steiglitz [15]) have been used for some time. In any event, in computation, once the designer has found the number of searches, N, required to satisfy his tolerance of error, if the criterion function possesses any regularity whatsoever, it would seem sensible to sample at evenly spaced grid points rather than randomly chosen points as per the preceding algorithms. We suspect that the procedures we have proposed may have merit if the function f is easily evaluated (such as in linear or quadratic programming problems, etc.) Regardless of its computational merits (or lack thereof), the preceding analysis should have practical value in pointing out that certain search problems which are much more difficult than those currently studied are, in principle at least, amenable to solution.

Our viewpoint and procedures differ from other approaches to the sequential search problem in that the nature of the domain space X can be suppressed. As noted above, the dimension of X plays little role, and in contrast with many other studies, the closeness of the operating point x to an optimizing point x* is of no consequence; it is on the closeness of f(x) to f(x*) that our attention focuses.

REFERENCES

1. Hadley, G. "Nonlinear and Dynamic Programming," Addison Wesley, Reading, Mass., 1964.

2. Spang, H.A. III. "A Review of Minimization Techniques for Nonlinear Functions," SIAM Review, 4, (1962), pp 343-365.

3. Kowalik, J. and M. Osborne. "Methods for Unconstrained Optimization Problems," Elsevier, New York, N.Y., 1969.

4. Kiefer, J. "Sequential Minimax Search for a Maximum," Proc. Amer. Math. Soc., 4, (1953 , pp 503-506.

5. Kiefer, J. "Optimum Sequential Search and Approximation Under Minimum Regularity Assumptions," SIAM Journal, 5, (1957), pp 105-136.

6. Bellman, R. and S. Dreyfus. "Applied Dynamic Programming," Princeton University Press, Princeton, N.J., 1962.

7. Kushner, H. "A Versatile Stochastic Model of a Function of Unknown and Time-varying Form," J. Math. Anal. and Appl. 5, (1962), pp 150-167.

8. Kushner, H. "A New Method for Locating the Maximum Point in an Arbitrary Multipeak Curve in the Presence of Noise," ASME J. Basic Engr., 86, (1964), pp 97-106.

9. Brooks, S. "A Discussion of Random Methods for Seeking Maximum," J. Opers. Res. Soc. of Amer., 6, (1958) pp 244-251.

10. Massey, F. "A Note on the Estimation of a Distribution Function by Confidence Limits," Ann. Math. Statist., 22, (1950), pp 116-119.

11. Robbins, H. and S. Monro. "A Stochastic Approximation Method," Ann. Math. Statist., 22, (1951), pp 400-407.

12. Kiefer, J. and J. Wolfowitz. "Stochastic Estimation of the Maximum of a Regression Function," Ann. Math. Statist., 23, (1952), pp 462-466.

13. Schmetterer, L. "Stochastic Approximation," Fourth Berkeley Symposium on Probability and Statistics, Vol. I, University of California Press, Berkeley, California, (1961), pp 587-609.

14. Kiefer, J. and J. Wolfowitz. "On the Deviations of the Empiric Distribution Function of Vector Chance Variables," *Trans. Amer. Math. Soc.*, *87*, (1958), pp 173-186.

15. Schumer, M. and K. Steiglitz. "Adaptive Step Size Random Search," *IEEE Trans. Auto. Control*, *13*, (1968), pp 270-276.